INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Combining simulation and mean-field like methods for inference in Hidden Markov Random Fields*

Florence Forbes — Gersende Fort

**N° 5721 – version 2**

version initiale Octobre 2005 – version révisée Janvier 2006

Thème COG


ISSN 0249-6399 ISRN INRIA/RR--5814--FR+ENG

INRIA
RHÔNE-ALPES

# Combining simulation and mean-field like methods for inference in Hidden Markov Random Fields

Florence Forbes*, Gersende Fort †

Thème COG — Systèmes cognitifs
Projet Mistis

Rapport de recherche n° 5721 – version 2‡— version initiale Octobre 2005 — version révisée Janvier 2006 37 pages

**Abstract:** Issues involving missing data are typical settings where exact inference is not tractable as soon as non trivial interactions occur between the missing variables. Approximations are required and most of them are based either on simulation methods or on deterministic variational methods. While variational methods provide fast and reasonable approximate estimates in many scenarios, simulation methods offer more consideration of important theoretical issues such as accuracy of the approximation and convergence of the algorithms but at a much higher computational cost. In this work, we propose a new class of algorithms that combine the main features and advantages of both simulation and deterministic methods and consider applications to inference in Hidden Markov Random Fields (HMRF). These algorithms can be viewed as stochastic perturbations of the Variational Expectation Maximization (VEM) algorithms, which are not tractable for HMRF. We focus more specifically on one of this perturbation and we prove its (almost sure) convergence to the same limit set as the limit set of VEM. In addition, experiments on synthetic and real-world images show that the algorithm performance is very close and sometimes better than that of other existing simulations-based and variational EM-like algorithms.

**Key-words:** Variational EM, Hidden Markov random fields, Monte-Carlo Markov Chain-based approximations, Image segmentation.

* MISTIS team, INRIA Rhône-Alpes
† LTCI, CNRS/ENST, 46 rue Barrault, 75634 Paris Cedex 13, France
‡ Changes from version1: the point of view and the positioning of this work has changed (see the new title and introduction). This led to additional experiments and new analysis and conclusions. Comparison with others algorithms has been added. Some of the proofs have been simplified.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

# Combiner simulations et méthodes en champ moyen pour l'inférence de champs de Markov cachés

**Résumé :** Les problèmes à données manquantes sont des cas typiques où l'inférence exacte n'est pas possible dès que l'on souhaite prendre en compte des dépendances non triviales entre variables cachées. Des approximations sont nécessaires et sont géneralement basées sur des méthodes de simulations ou sur des méthodes variationnelles déterministes. Les méthodes variationnelles fournissent des estimations rapides et raisonnables dans beaucoup de cas mais les méthodes à base de simulations, malgré leur coût calculatoire plus élevé, apportent bien souvent plus de réponses et de garanties sur des questions théoriques importantes telles que la qualité de l'approximation et la convergence des algorithmes. Dans ce travail, nous proposons une nouvelle classe d'algorithmes qui intègrent les caractéristiques et les avantages principaux de chacune des approches. En guise d'illustration, nous considérons une application à l'inférence des champs de Markov cachés et à la segmentation d'image. Ces algorithmes peuvent être vus comme des pertubations stochastiques des algorithmes EM variationnels (VEM) qui ne sont pas implementables dans le cas des champs de Markov cachés. Nous étudions plus précisément l'une de ces perturbations et nous prouvons sa convergence presque-sûre vers le même ensemble limite que celui de VEM. De plus, nous montrons sur des images synthétiques et réelles que les performances de ce nouvel algorithme sont très proches et parfois meilleures que celles d'autres algorithmes de type EM existant, à base de méthodes de simulations et/ou de méthodes variationnelles.

**Mots-clés :** EM variationnel, champs de Markov cachés, approximations par chaines de Markov Monte Carlo, segmentation d'images

# 1 Introduction

Missing data models are commonly used in various applications including areas as diverse as signal and image processing, genetics and epidemiology. They reveal very useful in modeling variability and heterogeneity in data and in solving various labeling or clustering issues. Due to the missing data structure, inference and parameter estimation tasks in such models often yield procedures that are intractable as soon as non trivial interactions in the data are taken into account. In most applications, their complexity requires the development of approximations techniques. These techniques are usually based either on deterministic numerical methods such as variational methods (*e.g.* [22, 38]) or on simulation methods such as Monte Carlo Markov Chains (MCMC) techniques (*e.g.* [32]). Choosing one or the other approach can be advantageous depending on the context and the goal in mind. Inference problems are usually formulated as the computation of a quantity of interest (*e.g.* a probability distribution) as the solution to an optimization problem. Variational inference methods refer to a certain class of deterministic methods that consist of solving a perturbed version of the optimization problem. In order to define such a problem, it is necessary to specify both a cost function to be optimized and a constraint set over which the optimization takes place. Variational methods then arise as *relaxations*, that is, simplified optimization problems that involve some approximation of the constraint set, the cost function or both. The original issue is replaced by an easier optimization problem and variational methods have been shown to provide fast and reasonable approximate estimates in many scenarios [22]. However, since it is often the case that there are no other feasible choices rather than to resort to variational approximation in practical situations, it appears frequently that these approximations are being used to practical problems with little consideration of important issues such as accuracy of the approximation, convergence of the algorithms and so on. Convergence results exist for the so-called Variational Expectation Maximization (VEM) algorithms (see [4]) but their application is restricted to specific settings which limit the kind of interactions allowed between the missing data to very simple ones. Variants to extend the application domain of algorithms such as VEM have been proposed (see *e.g.* [45] and [6] in an image segmentation framework) but they did not succeed in preserving the convergence results. As a matter of fact, in most settings of practical interest, theoretical results regarding accuracy and convergence properties are still missing. Simulation methods appear then as natural candidates to make algorithms tractable for a wider class of problems while providing tools to study their convergence. As an example, convergence of MCMC based algorithms have been widely studied and a lot of tools are now available that make various convergence results available or at least easy to derive (see for instance [15] for a convergence proof of the Monte-Carlo EM algorithm of [39] based on Monte Carlo integration procedure with MCMC sampling techniques). In this paper, our aim is to show that combining both type of methods to design new algorithms can greatly improve accuracy and modeling flexibility in missing data settings. The main idea is that algorithms resulting from such a combination will benefit from the good features of both approaches simultaneously. Deterministic schemes are easy to implement and can provide fast estimates

while simulation methods often lead to more accurate results with guaranteed convergence. There have been other attempts at combining approximation techniques and simulation methods. The closest in spirit to our approach is that in [9]. The authors introduce a class of MCMC algorithms that use variational approximations as initial proposal distributions and consider an application to sigmoidal belief networks. In our work, we use a different approach and different tools. We rather incorporate MCMC simulations into variational algorithms and focus on a different application. Other attempts in the statistics community include the use of Laplace approximation with simulation techniques [21], the Gibbsian-EM [7], the Restoration-Maximization algorithm [30], the Monte-Carlo approximations by [31] and more recently, the Simulated Field algorithm of [6]. However, most of these procedures were not originally designed with this combining idea in mind and no convergence results are available for them. A detailed comparison of some of these algorithms, for the case of hidden binary isotropic Markov chains, can be found in [2].

Image segmentation and Hidden Markov Random Fields (HMRF) estimation is a typical setting where one encounters these tradeoffs between accuracy, convergence guarantees and reasonable computing time. Difficulties arise due to the dependence structure in the models. The Expectation Maximization (EM) algorithm [11], typically used in missing data cases, yields update procedures that do not have a closed form expression and is intractable analytically. Different algorithms have been proposed to overcome this intractability of EM. Among *pure* simulation techniques, a straightforward variant of the Monte-Carlo EM algorithm can be used (see Section 5) while variational versions of EM are deterministic alternatives. In particular, VEM algorithms have been popular in cases where the E-step of EM is intractable [22]. The most popular class of VEM procedures is certainly the mean-field EM one. The mean field approach consists in computing quantities related to a complex probability distribution, by using a simple tractable model such as the family of independent distributions. However, introducing relaxation in the E-step does not fully answer the question of inference in cases where the M-step remains intractable due to the complex structure of dependence between the hidden variables. It follows that VEM algorithms cannot be directly applied in the HMRF segmentation framework where additional approximations are required in the M-step. Further algorithms have then been designed, that propagate the relaxation in the E-step to the M-step. The combination in such a way of the mean field theory and the EM procedures for HMRF is due to [45]. Using ideas from this principle, [6] proposed, in the context of Markovian image segmentation, a class of EM-like algorithms generalizing [45] which show good performance in practice. In this work, we present another way to overcome the intractability of VEM based on the idea of combining deterministic and simulation-based approximations. We start from VEM procedures for which convergence properties are well established and introduce simulations in these algorithms. In addition to make the algorithms tractable, we claim that the introduction of a small perturbation at each iteration of VEM, yields algorithms with the same asymptotic behavior as VEM. More specifically, we propose a class of (stochastically) perturbed VEM algorithms where the noise at each iteration is controlled so that it gets negligible, in a sense to be specified, when the number of iterations tends to infinity. We prove our claim by adapting the results of [15] relative to perturbed iterative maps. We propose an example of such a stochastic VEM algorithm, the Monte-Carlo VEM algorithm (MCVEM) which is tractable in practice and for which we prove convergence results. In addition, the algorithm performance is compared, on synthetic and real-world images, with various other algorithms that are typical of one of the approach separately. For deterministic algorithms, we report the comparison with the Mean Field algorithm of [6] while for pure MCMC techniques, we consider a simple extension of the MCEM algorithm, the later being intractable in the HMRF setting. As an illustration, we also compare with two other algorithms among the ones that combine simulation and deterministic methods, namely the Gibbsian-EM and the Simulated Field algorithms, chosen for their flexibility in missing data problems. We

observe that the MCVEM algorithm provides the best or very close to the best results for most of our test images. Our algorithm has thus many advantages: (a) it is tractable in practice, (b) we are able to prove convergence results so that the set of its limit points is identified (as being the set of the limit points of VEM), and (c) it is efficient when applied to image segmentation. It illustrates how combining deterministic and simulation techniques can result in improved algorithms.

The paper is organized as follows. In section 2, we first state the Markov model-based image segmentation problem and present the variational principle through a description of the VEM algorithm. The combination with MCMC simulations is specified in section 3 through the presentation of our MCVEM algorithm. Its links to other EM variants are also specified in this section. Convergence theorems are given in section 4 with their proofs postponed in Appendix. We show that these theorems apply for our image segmentation purpose and report experiments on synthetic and real-world images in section 5 illustrating the good performance of the algorithm. A discussion section concludes the paper.

# 2 Markov model-based image segmentation

Problems involving incomplete data, where part of the data is missing or unobservable, are common in image analysis. The aim may be to recover an original image which is hidden and has to be estimated from a noisy or blurred version. More generally, the observed and hidden data are not necessarily of the same nature. The observations may represent measurements, *e.g.* multidimensional variables recorded for each pixel of an image while the hidden data could consist of an unknown class assignment to be estimated, for each pixel, from the observations. This case is usually referred to as image segmentation. In this paper, we focus on Markov model-based image segmentation. In Section 2.1, we recall basic definitions concerning the Markov models used for the unobserved data and specify the complete parametric models for the observed and unobserved data. We recall the EM algorithm in section 2.2. Its extension to VEM, which is the basis of our parameter estimation algorithm, is specified in Section 2.3.

## 2.1 Hidden Markov random fields for segmentation

Let $S$ be a finite set of sites with a neighborhood system defined on it. Let $N = |S|$ denote the number of sites. A typical example in image analysis is the two dimensional lattice with a first order neighborhood system: for each site, the neighbors are the four sites surrounding it. A set of sites $C$ is called a clique if it contains sites that are all neighbors. Let $V$ be a finite set with $K$ elements. Each of them will be represented by a binary vector of length $K$ with one component being 1, all others being 0, so that $V$ will be seen as included in $\{0,1\}^{K\times K}$ and its elements denoted by $\{e_1,\ldots,e_K\}$. We define a discrete Markov random field as a collection of discrete random variables, $\mathbf{Z} = \{Z_i,\ i \in S\}$, defined on $S$, each $Z_i$ taking values in $V$, whose joint probability distribution $p_Z$ satisfies the following properties,

$$\forall \mathbf{z},\quad p_Z(z_i \mid \mathbf{z}_{S\setminus\{i\}}) = p_Z(z_i \mid z_j,\ j \in \mathcal{N}(i)) \tag{1}$$

$$\forall \mathbf{z},\quad p_Z(\mathbf{z}) > 0, \tag{2}$$

where $\mathbf{z}_{S\setminus\{i\}}$ denotes a realization of the field restricted to $S\setminus\{i\} = \{j \in S, j \neq i\}$ and $\mathcal{N}(i)$ denotes the set of neighbors of $i$. More generally, if $A$ is a subset of $S$, we will write $\mathbf{z}_A$

for $\{z_i,\ i \in A\}$. In words, property (1) means that the interactions between site $i$ and the other sites actually reduce to interactions with its neighbors. Property (2) is important for the Hammersley-Clifford theorem to hold. This theorem states that the joint probability distribution of a Markov field is a Gibbs distribution given by

$$p_Z(\mathbf{z}) \quad = \quad W^{-1} \exp(-H(\mathbf{z})), \tag{3}$$

where $H$ is the energy function

$$H(\mathbf{z}) \quad = \quad \sum_c V_c(\mathbf{z}_c). \tag{4}$$

The $V_c$'s are the clique potentials and may depend on parameters, not specified in the notation. $W = \sum_{\mathbf{z}} \exp(-H(\mathbf{z}))$ is the normalizing factor also called the partition function; $\sum_{\mathbf{z}}$ (resp. $\sum_{\mathbf{z}_A}$) denotes a sum over all possible values of $\mathbf{z}$ (resp. $\mathbf{z}_A$). The computation of $W$ involves all possible realizations $\mathbf{z}$ of the Markov field. Therefore, it is, in general, exponentially complex, and not computationally feasible. This can be an issue when using these models in situations where an expression of the joint distribution $p_Z(\mathbf{z})$ is required. In the following sections, we will deal with approximation of $p_Z(\mathbf{z})$. We will denote by $\mathcal{Z} = V^N$ the set in which $\mathbf{Z}$ takes values and by $\mathcal{D}$ the set of probability distributions on $\mathcal{Z}$.

Image segmentation involves observed variables and unobserved variables which have to be recovered. The hidden variables are modeled as a discrete Markov random field, $\mathbf{Z}$, with distribution $p_Z$ as defined in (3) and an energy function $H$ depending on a parameter $\beta \in \mathcal{B} \subseteq \mathbb{R}$ and henceforth denoted by $H(\mathbf{z};\beta)$. It is assumed that the observations $\mathbf{Y}$ are conditionally independent given the Markov random field $\mathbf{Z}$, with conditional distribution $p_{Y|Z}$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$, where $n_\theta$ is the dimension of $\theta$ depending on the model under consideration.
In the general case, the likelihood of $(\mathbf{Y}, \mathbf{Z})$ called the complete likelihood $p_{(Y,Z)}$ is given by

$$\begin{aligned} p_{(Y,Z)}(\mathbf{y}, \mathbf{z}\ ;\ \theta, \beta) \quad &= \quad p_{Y|Z}(\mathbf{y} \mid \mathbf{z}; \theta)\ p_Z(\mathbf{z}; \beta) \\ &= \quad W(\beta)^{-1} \exp\{-H(\mathbf{z};\beta) + \log p_{Y|Z}(\mathbf{y} \mid \mathbf{z}; \theta)\}. \end{aligned} \tag{5}$$

Then the conditional likelihood of the hidden variables $Z$ given the observations $Y$, $p_{Z|Y}$ is given by $p_{Z|Y} = p_{(Y,Z)}/p_Y$ where $p_Y$ is the likelihood of the observations $Y$ (called the *incomplete* likelihood). It is easy to see from (5) that the conditional field $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is a Markov field as $\mathbf{Z}$ is. Its energy function is

$$H(\mathbf{z} \mid \mathbf{y}; \theta, \beta) \ = \ H(\mathbf{z}; \beta) - \log p_{Y|Z}(\mathbf{y} \mid \mathbf{z}; \theta).$$

In the following developments, we will refer to Markov fields $\mathbf{Z}$ and $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ as the marginal and conditional fields.
In image segmentation problems, the question of interest is generally to recover the unknown image $\mathbf{z}$, interpreted as a classification into a finite number $K$ of labels. This classification usually requires values for the vector parameter $\psi = (\theta, \beta)$.

## 2.2 Inference by EM algorithm

If unknown, the parameters are usually estimated in the maximum likelihood sense

$$\hat{\psi} = \text{argmax}_{\psi \in \Psi} \ln p_Y(\mathbf{y}; \psi), \tag{6}$$

where $\Psi = \Theta \times \mathcal{B}$ is the parameter space. This optimization is usually solved by the iterative EM procedure [11]. Any iteration may be formally decomposed into two steps: given the current value of the parameter $\psi^t$, the so-called E-step consists in restoring the missing data, *i.e.* in computing

$$\mathcal{Q}(\psi; \psi^t) = \sum_{\mathbf{z} \in \mathcal{Z}} \ln p_{(Y,Z)}(\mathbf{y}, \mathbf{z}; \psi)\ p_{Z|Y}(\mathbf{z}|\mathbf{y}; \psi^t),$$

the expectation of the complete log-likelihood knowing the observations $\mathbf{y}$ and the current estimate $\psi^t$. The parameter is then updated and defined as maximizing this expected complete log-likelihood

$$\psi^{t+1} = \text{argmax}_{\psi \in \Psi}\, \mathcal{Q}(\psi; \psi^t). \tag{7}$$

It is known that, under regularity conditions, EM converges to the set of the stationary points of the incomplete likelihood $\psi \mapsto p_Y(\mathbf{y}; \psi)$ [41]. As discussed in [8] and [27], EM can be viewed as an alternating maximization procedure of a function $F$ defined below (equation (9)), by rewriting the objective function $\ln p_Y(\mathbf{y}; \cdot)$ as follows: for any probability distribution $q \in \mathcal{D}$, it holds

$$\ln p_Y(\mathbf{y}; \psi) = F(q, \psi) + \text{KL}\left(q; p_{Z|Y}(\cdot|\mathbf{y}; \psi)\right) \tag{8}$$

where

$$F(q, \psi) = \sum_{\mathbf{z} \in \mathcal{Z}} \ln \left( \frac{p_{(Y,Z)}(\mathbf{y}, \mathbf{z}; \psi)}{q(\mathbf{z})} \right)\ q(\mathbf{z}), \tag{9}$$

and KL denotes the Kullback-Leibler divergence,

$$\forall p_1, p_2 \in \mathcal{D}, \quad \text{KL}(p_1; p_2) = \sum_{\mathbf{z} \in \mathcal{Z}} \ln \left( \frac{p_1(\mathbf{z})}{p_2(\mathbf{z})} \right)\ p_1(\mathbf{z}). \tag{10}$$

The Kullback-Leibler divergence is a measure of dissimilarity between two distributions. It is always positive and is zero when and only when the distributions are equal. The alternating maximization algorithm is an iterative procedure; for a current value $(q^t, \psi^t) \in \mathcal{D} \times \Psi$, set

$$\begin{aligned} q^{t+1} &= \text{argmin}_{q \in \mathcal{D}}\ \text{KL}\left(q; p_{Z|Y}(\cdot|\mathbf{y}; \psi^t)\right) \\ &= \text{argmax}_{q \in \mathcal{D}}\ F(q, \psi^t), \end{aligned} \tag{11}$$

and

$$\begin{aligned} \psi^{t+1} &= \text{argmax}_{\psi \in \Psi}\ F(q^{t+1}, \psi) \\ &= \text{argmax}_{\psi \in \Psi} \sum_{\mathbf{z} \in \mathcal{Z}} \ln p_{(Y,Z)}(\mathbf{y}, \mathbf{z}; \psi)\ q^{t+1}(\mathbf{z}). \end{aligned} \tag{12}$$

Observe that the maximization (12) can also be understood as the minimization of a Kullback-Leibler divergence, up to some convention on $p_Y$ thus justifying the name of alternating minimization procedure often found in the literature (e.g. [8, 4]). Roughly speaking, given $(q^t, \psi^t)$, the algorithm consists in finding $q^{t+1}$ such that the error when approximating $\ln p_Y(\mathbf{y}; \psi^t)$ by $F(q, \psi^t)$ is minimal. For this optimal $q^{t+1}$, it then finds $\psi$ such that the minorizing bound $F(q^{t+1}, \psi)$ of the objective function $\psi \mapsto \ln p_Y(\mathbf{y}; \cdot)$ is maximal. The first optimization (11) has an explicit solution $q^{t+1} = p_{Z|Y}(\cdot|\mathbf{y}; \psi^t)$. Hence, according to (12), $\text{argmax}_{\psi \in \Psi}\ F(q^{t+1}, \psi) = \text{argmax}_{\psi \in \Psi}\ \mathcal{Q}(\psi, \psi^t)$ for all $t \geq 0$ and the "marginal" sequence $\{\psi^t\}_t$ of the sequence $\{(q^t, \psi^t)\}_t$ produced by the alternating maximization procedure is an EM path.

There exist different generalizations of EM when the M-step (7) is intractable; it can be relaxed by requiring just an increase rather than an optimum. This yields Generalized EM (GEM) procedures ([24]; see also [3] for a convergence result).

## 2.3 Inference by VEM algorithm

Unfortunately, EM (or GEM) is not appropriate for solving the optimization problem (6) in Hidden Markov Random Field due to the complex structure of the hidden variables $Z$; in practice, the distribution $p_Z(\mathbf{z};\beta)$ is only known up to a multiplicative constant (*i.e.* up to the partition function) and the domain $\mathcal{Z}$ is too large so that the E-step is intractable. Alternative approaches were proposed and they can be understood as generalizations of the alternating maximization procedures developed above. The optimization (11) can be solved over a restricted class of probability distribution $\tilde{\mathcal{D}}$ on $\mathcal{Z}$. The quantity $F(q^{t+1},\psi)$ is computed with some optimal distribution $q^{t+1} \in \tilde{\mathcal{D}}$ and the M-step (12) remains unchanged. This yields the Variational EM (VEM) algorithms [22]. VEM can also be introduced as resulting from a relaxation of a convex optimization problem; the objective function $p_Y(\mathbf{y};\cdot)$ is re-written as the ratio of two partition functions and VEM results from the approximation of one of them using the notion of conjugate duality in convex analysis (see [37] and [38] for details).
[4] proved that, under mild regularity conditions, VEM converges to the set $\mathcal{L}$ of the stationary points of the function $F$ in $\tilde{\mathcal{D}}$. Here again, generalizations of VEM can be defined by requiring an increase rather that an optimum in the M-step (12) thus defining generalized VEM procedures. These relaxation methods are part of the Generalized Alternating Minimization procedures [4].

The most popular form of VEM is the case when $\tilde{\mathcal{D}}$ is the set of the independent probability distributions on $\mathcal{Z}$ so that $q^{t+1}(\mathbf{z})$ is a factorized distribution $\prod_{i\in\mathcal{S}} q_i^{t+1}(z_i)$. Computing the gradient of the Kullback-Leibler divergence with regards to $q_i^{t+1}(e_k)$, $i \in S$ and $e_k \in V$, and setting it to zero, leads to a fixed point equation:

$$\forall i \in S, \forall e_k \in V, \quad \ln q_i^{t+1}(e_k) \quad = \quad c_i + \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi^t)\ \{\delta_{e_k}(z_i)\ \prod_{j\neq i} q_j^{t+1}(z_j)\} \quad (13)$$

where $c_i$ is the normalizing constant and $\delta_e$ denotes the Dirac mass at point $e$. The Markov property implies that the right-hand side of the equation only involves the probability distributions $q_j$, $j \in \mathcal{N}(i)$. This equation can also be recovered from a different point of view. The idea when considering a particular site $i$ is to neglect the fluctuations of the sites interacting with $i$ so that the resulting system behaves as one composed of independent variables. More specifically, for all $j$ different from $i$, the $Z_j$'s are fixed to their current conditional mean value $E[Z_j|\mathbf{y};\psi^t]$. However, these mean values are unknown and it is originally the goal of the approximation to compute them. Therefore, the approximation depends on a self-consistency condition which is that the mean values that can be computed from the approximate distribution must be equal to the mean values used to define this approximate distribution. Then, replacing the exact conditional mean values by the mean values in the approximation leads to a fixed point equation involving these mean values (see [6] for more details). Existence and uniqueness of a solution to (13) are properties that have not yet been fully understood and will not be discussed here. We refer to [36] for a better insight into the properties of the (potentially multiple) solutions of the mean field equations. Such solutions are usually computed iteratively (see [40] and [1], [46] and an erratum [13]). We will discuss in Section 5 the consequences of the non-unicity of the solution when running mean-field based procedures for image segmentation.

Despite the relaxation which may make the summation of the VEM E-step explicit for a convenient choice of $\tilde{\mathcal{D}}$ *i.e.* the computation of $F(q^{t+1},\psi)$ in (12), VEM remains intractable for hidden Markov random fields. Indeed, for such models, the M-step (12) decomposes, omitting the dependence in $q^{t+1}$ in the notation, into $\theta^{t+1} = \text{argmax}_{\theta\in\Theta} \mathcal{Q}_1^{t+1}(\theta)$

and $\beta^{t+1} = \text{argmax}_{\beta \in \mathcal{B}} \mathcal{Q}_2^{t+1}(\beta)$ where

$$\mathcal{Q}_1^{t+1}(\theta) = \sum_{\mathbf{z} \in \mathcal{Z}} \ln p_{Y|Z}(\mathbf{y}|\mathbf{z};\theta)\ q^{t+1}(\mathbf{z}), \tag{14}$$

$$\mathcal{Q}_2^{t+1}(\beta) = \sum_{\mathbf{z} \in \mathcal{Z}} \ln p_Z(\mathbf{z};\beta)\ q^{t+1}(\mathbf{z}), \tag{15}$$

$$= -\sum_{\mathbf{z} \in \mathcal{Z}} H(\mathbf{z};\beta) q^{t+1}(\mathbf{z}) - \ln W(\beta). \tag{16}$$

Under additional commonly used assumptions on $p_{Y|Z}$, equation (14) can usually be solved in closed form (see for example section 5). The issue is to solve equation (16) since the update of $\beta$ requires an explicit expression of the partition function or an explicit expression of some related quantities (its gradient for example).
To overcome this difficulty, different approaches were proposed. We review in Section 3.1 some of these methods based on mean-field relaxations of EM: the relaxation applied in the E-step is then propagated to the M-step. The theoretical contribution of this paper states that introducing noise at each VEM iteration in such a way that this perturbation gets to zero (in a sense to be specified) as the number of iterations increases, yields an algorithm which has the same asymptotic behavior as VEM. This noise is defined in order to make VEM tractable for solving inference in hidden MRF. We thus propose in Section 3.2 an example of such procedures: our stochastically perturbed version of VEM consists in approximating the partition function $W(\beta)$ by some Monte-Carlo sum.

# 3 Variational EM-like algorithms

In the first approach we consider here (section 3.1), the procedures can be viewed as the standard EM algorithm applied with a different independent mixture model at each iteration, corresponding to a simplified distribution. In section 3.2, the approach differs in that the approximation method does not lead to a simple valid model but appears as a succession of approximations to overcome successive computational difficulties.

Let $\mathcal{I}$ be the set of independent probability distributions on $\mathcal{Z}$ and $\mathcal{I}_r$ be the set of independent probability distributions on $\mathcal{Z}$ such that $q \in \mathcal{I}_r$ implies that $\forall e_k \in V$, $\sum_{i=1}^{N} q_i(e_k) \neq 0$; $\mathcal{I}_r$ contains the independent probability distributions on $\mathcal{Z}$ such that the probability that no pixels are labeled $k$ is zero.

## 3.1 Mean field and Simulated field algorithms

The algorithms proposed in [6] are alternatives to VEM that propagate the approximation $q^{t+1}$ of $p_{Z|Y}(\mathbf{z}|\mathbf{y},\psi^t)$ to $p_Z(\mathbf{z};\beta)$. They are based on the observation that both $p_Z(\mathbf{z};\beta)$ and $p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi)$ are not available but $p_{Y|Z}(\mathbf{y}|\mathbf{z};\theta)$ is (see an illustration in section 5). Furthermore, knowing $p_{Y|Z}(\mathbf{y}|\mathbf{z};\theta)$, it is enough to approximate one of the unknown quantities, either $p_Z(\mathbf{z};\beta)$ or $p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi)$, to derive an approximation of the other and of the joint distribution $p_{(Y,Z)}$. The authors in [6] use then for $p_Z$ an approximation obtained by applying the Bayes formula to the exact $p_{Y|Z}$ and to $q^{t+1}$ which approximates $p_{Z|Y}(\mathbf{z}|\mathbf{y};\psi^t)$. It follows an approximation of $p_Z$ that factorizes so that all the terms in (15) become easy to compute. When $q^{t+1}$ is obtained by solving the mean field equations (13), this results in the so-called *Mean Field algorithm* of [45] but other variants are investigated in [6] based

on other factorized distributions. In particular, when the neighbors at each site are set to simulations instead of mean values, the algorithm becomes stochastic. The *Simulated Field algorithm* described in [6] follows this idea: the neighbors are drawn at random and set to the realization, after one iteration, of a Markov chain with stationary distribution $p_{Z|Y}(\cdot; \mathbf{y}, \psi^t)$. Experiments in [6] and [14] show that better performance is obtained with this latter algorithm.
An important implementation detail is related to the global optimization strategy adopted in [6]. The fixed point equation (13) defines a functional relationship between its left and right-hand sides. The equation is then usually solved iteratively by applying the functional relationship until $q^{t+1}$ no further moves. In [6], the idea is to immediately take into account each change in $q^{t+1}$ when updating the other parameters $(\theta, \beta)$. It follows that the factorized distribution $q^{t+1}$ does not necessarily solve (13) (in the *Mean Field* case) but is the result of only one iteration. The *Simulated Field algorithm* is implemented according to the same strategy.

## 3.2 The Monte-Carlo VEM algorithm

The Monte-Carlo VEM (MCVEM) algorithm illustrates a novel strategy that combines MCMC and variational methods. The algorithm consists in running a particular VEM procedure (namely the one corresponding to equation (13)), and replacing the update of the parameter $\beta$ by the maximization of some quantity in which the partition function $W(\beta)$ is approximated by a Monte-Carlo sum. This yields the following iterative procedure.
Fix a positive sequence $\{\gamma^t\}_t$ such that $\inf_t \gamma^t > 0$, a sequence of positive integers $\{J_t\}_t$ and a sequence of probability distributions $\{\pi^t\}_t$ on $\mathcal{Z}$. For the current value $(q^t, \psi^t)$ of the parameter:

(i) Update the $q$-component

$$q^{t+1} = \text{argmin}_{q \in \mathcal{I}_r} \text{KL}\left(q; p_{Z|Y}(\cdot|\mathbf{y}; \psi^t)\right).$$

(ii) Sample a Markov random field $\{Z^{j,t}\}_{1 \le j \le J_t}$ with invariant distribution $\pi^t(\mathbf{z})$ and compute $\tilde{\mathcal{Q}}_2^{t+1}(\beta) = -\{\sum_{\mathbf{z} \in \mathcal{Z}} H(\mathbf{z}; \beta) q^{t+1}(\mathbf{z}) + \ln \tilde{W}^{J_t, \pi^t}(\beta)\}$ where

$$\tilde{W}^{J_t, \pi^t}(\beta) = \frac{1}{J_t} \sum_{j=1}^{J_t} \exp\left(-H(\mathbf{z}^{j,t}; \beta) - \ln \pi^t(\mathbf{z}^{j,t})\right).$$

(iii) Update the $\psi$-component by setting $\theta^{t+1} = \text{argmax}_{\theta \in \Theta}\ \mathcal{Q}_1^{t+1}(\theta)$ and $\beta^{t+1} = \text{argmax}_{\beta \in \mathcal{B}^t}\ \tilde{\mathcal{Q}}_2^{t+1}(\beta)$,

where $\mathcal{B}^t = \{\beta \in \mathcal{B}, |\beta - \beta^t| \le \gamma^t\}$.
In practice, the step (i) is implemented directly by iteratively solving a nonlinear system given by the mean field equations (13).
The step (ii) followed by the update of the $\beta$-component looks like the algorithm proposed by [20] for maximum likelihood parameter estimation of exponential families, except that, in [20], it is assumed that the samples are independent and identically distributed. In MCVEM, the rough idea is that the partition function is approximated by a Monte-Carlo sampling from some distribution $\pi^t$, thus using an importance sampling estimator (or possibly a self-normalized importance sampling estimator [19]). If the sampler is good enough so that a

law of large numbers holds, $\lim_{J\to\infty} \tilde{W}^{J,\pi^t}(\beta) = W(\beta)$, and one can expect that by choosing $J$ large enough, $\tilde{W}^{J,\pi^t}(\beta)$ provides a good approximation of $W(\beta)$. As discussed in [43], the best choice for approximating $W(\beta)$ is $\pi = p_Z(\cdot;\beta)$. This is useless for our purposes since we want a good approximation of $W(\beta)$ whatever $\beta$, for a given distribution $\pi$. Nevertheless, by choosing $\pi^t = p_Z(\cdot;\beta^t)$, it can be expected that for some $J$ sufficiently large, $\tilde{W}^{J,\pi^t}(\beta)$ is a good approximation of the partition function $W(\beta)$ in a neighborhood of $\beta^t$. This explains the local optimization of $\tilde{\mathcal{Q}}_2^{t+1}(\beta)$ and the introduction of the domain $\mathcal{B}^t$. We assumed throughout this paper that $\{\gamma^t\}_t$ is a deterministic sequence uniformly bounded away from zero. One could be interested in choosing $\gamma^t$ as a function of the Hessian of $\tilde{Q}_2^{t+1}(\beta^t)$; in that case, $\{\gamma^t\}_t$ is a random sequence and the study of the asymptotic behavior of MCVEM is slightly more complex. We will discuss this extension in Section 4.5.
In practice, we choose $\pi^t = p_Z(\cdot;\beta^t)$ and sample the Markov random field by using Monte-Carlo Markov Chain samplers (Gibbs sampler [18], Hastings-Metropolis sampler [25], Swendsen-Wang sampler [35], $\cdots$). Observe that $\tilde{W}^{J,\pi^t}(\beta)$ can be known up to a multiplicative constant independent of $\beta$: this allows the choice $\pi^t = p_Z(\cdot;\beta^t)$ which is known up to the partition function $W(\beta^t)$. Here after, we will simply write $\tilde{W}^{J,\beta^t}$ as a shorthand notation for $\tilde{W}^{J,p_Z(\cdot;\beta^t)}$.

We have found this simple method for estimating $W(\beta)$ easy to use and very satisfying in our experiments that we chose as typical segmentation problems (see Section 5). However, we are aware of possible limitations of such MCMC samplers. In practice our numerical results could certainly be further improved by using more sophisticated methods. A full analysis of the problem of estimating normalizing constants has been given by [17]. They discussed several methods that are more sophisticated but also more cumbersome. In this paper, our focus is mainly on convergence results and on showing that it is advantageous to combine variational and MCMC methods. We did not investigate further the possibility of using better samplers.

Due to the simulation step, MCVEM is a stochastic algorithm. A difficulty, when dealing with random sequences $\{(q^t,\psi^t)\}_t$ is to guarantee the almost-sure boundedness. This can be done by a simple modification of the iterative scheme MCVEM, as described in [15] for the stabilization of the Monte-Carlo EM: let $\{\mathcal{C}^t\}_t$ be a sequence of compact subsets such that for any $t \geq 0$,

$$\mathcal{C}^t \subsetneq \mathcal{C}^{t+1} \qquad\qquad \mathcal{I}_r \times \Psi = \bigcup_{t\geq 0} \mathcal{C}_t. \tag{17}$$

Roughly speaking, it consists in introducing a variable $\tau^t$ which counts the number of reprojections from time 0 to time $t$ ($\tau^0 = 0$). At iteration $t+1$, the candidate $(q^{t+1},\psi^{t+1})$ has to be in the compact $\mathcal{C}^{\tau^t}$; otherwise, the sequence $\{(q^t,\psi^t)\}_t$ is reinitialized: $(q^{t+1},\psi^{t+1})$ is replaced by $(q^0,\psi^0)$, and $\tau^{t+1} = \tau^t + 1$. A detailed algorithmic description can be found in [15].

# 4 Convergence theorems for stochastically perturbed VEM algorithms

We first address the convergence of a generalized VEM algorithm: in this generalization, the update of the $\beta$-component is done by local optimization on the domain $\mathcal{B}^t$ of the function $\beta \mapsto \mathcal{Q}_2^{t+1}(\beta)$, while VEM requires a global optimization. We show that the key property

for convergence is the existence of a positive Lyapunov function (see Appendix 8 for a definition), namely the function $L = \exp(F)$ where $F$ is given by (9). Unfortunately, due to the introduction of a perturbation at every iteration of the (generalized) VEM algorithm, this function is not a Lyapunov function for MCVEM. To overcome this drawback, a result of interest is the following limit: for any compact set $\mathcal{K}$ in $\mathcal{I}_r \times \Psi$,

$$\lim_t \left| L(q^{t+1}, \psi^{t+1}) - L(q^{t+1}, \bar{\psi}^{t+1}) \right| \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} = 0, \tag{18}$$

almost surely when the perturbation is stochastic, where $(q^{t+1}, \psi^{t+1})$ results from one iteration of the perturbed algorithm and $(q^{t+1}, \bar{\psi}^{t+1})$ results from one iteration of generalized VEM when both the algorithms are started at $(q^t, \psi^t)$. [15] proved that when this limit holds, the perturbed algorithm and the exact one have the same asymptotic behavior (under additional regularity conditions omitted at this level of the exposition). Condition (18) means that the perturbation vanishes along the compact path, when measured in terms of the error induced on the Lyapunov function; as a consequence, the perturbed algorithm inherits the effect of the Lyapunov function and converges to the same limit set as the original (unperturbed) generalized VEM algorithm. This sufficient condition is adapted from results on perturbed iterative random maps [15]. When applied to MCVEM, we will show that this condition is nothing more than a condition on the Monte-Carlo approximation: if the fluctuations of the Monte-Carlo approximation of $W(\beta)$ by $\tilde{W}^{J_t, \beta^t}$ are well enough controlled (in a sense to be specified) and the number of simulations $J_t$ goes to $+\infty$ when $t \to +\infty$ at a rate depending upon the control, then MCVEM and VEM have the same asymptotic behavior.
Condition (18) requires a compact path : we extend the stabilization procedure in [15] to the present framework. When applied to MCVEM, this yields the re-projection algorithm derived in Section 3.2; we show that the number of re-projections is finite almost-surely, so that the stable MCVEM path remains compact and stable MCVEM and MCVEM have the same asymptotic behavior. This is our second result of convergence.
While the condition (18) applies to any (stochastic) perturbation of generalized VEM, we only consider MCVEM for clarity: we derive sufficient conditions on the model and on the approximation ensuring convergence of stable MCVEM. We detail the proofs and give the extension of the results by [15] on iterative random maps in Appendix 9, so that convergence results for any other perturbation of VEM can easily be deduced from the present work.

This section is organized as follows: we start with formulating sufficient conditions on the model (paragraph 4.1) and on the Monte-Carlo approximations (paragraph 4.2). We then state a theorem on the convergence of the generalized VEM algorithm (paragraph 4.3) and a theorem on the convergence of the MCVEM algorithm (paragraph 4.4). The central result of this section is Theorem 2 that addresses the convergence of the MCVEM paths. It is the main original theoretical contribution of this paper. The proofs of these theorems are postponed in Appendix 8 and 9. Possible extensions are briefly discussed in section 4.5.

## 4.1 Model assumptions

We assume that

**A1** $\mathcal{Z}$ is finite, $\mathcal{B} \subseteq \mathbb{R}$, $\Theta \subseteq \mathbb{R}^{n_\theta}$ and $\Psi = \Theta \times \mathcal{B}$.

**A2** (i) The function $\psi \mapsto p_{(Y,Z)}(\mathbf{y}, \mathbf{z}; \psi)$ is continuous on $\Psi$.
(ii) For all $q \in \mathcal{I}_r$, the set $\operatorname{argmax}_{\theta \in \Theta} \sum_{\mathbf{z} \in \mathcal{Z}} p_{Y|Z}(\mathbf{y}|\mathbf{z}; \theta) q(\mathbf{z})$ is not empty.

(iii) For any $\mathbf{z} \in \mathcal{Z}$, $\beta \mapsto H(z;\beta)$ is twice-continuously differentiable on $\mathcal{B}$. The function $\beta \mapsto -\{\sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z};\beta)q(\mathbf{z}) + \ln W(\beta)\}$ is strictly concave and admits a unique maximum in $\mathcal{B}$ for any $q \in \mathcal{I}$.

(iv) The function $\beta \mapsto -\{\sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z};\beta)q(\mathbf{z}) + \ln \tilde{W}^{J,b}(\beta)\}$ is strictly concave and admits a unique maximum in $\mathcal{B}$, for any $q \in \mathcal{I}_r$, any integer $J$ and any $b \in \mathcal{B}$.

Define $L$ on $\mathcal{I} \times \Psi$,

$$L(q,\psi) = \exp(F(q,\psi)), \tag{19}$$

where $F$ is given by (9).

**A3** For any $M > 0$, the level set $\{(q,\psi) \in \mathcal{I}_r \times \Psi, L(q,\psi) \geq M\}$ is bounded.

Under **A1** and **A2(i)**, $L$ is a continuous function and the level set is a closed subset of $\mathcal{I}_r \times \Psi$. Hence, it is compact in $\mathcal{I}_r \times \Psi$. Define

$$\mathcal{L} = \left\{(q^*,\psi^*) \in \mathcal{I}_r \times \Psi, q^* \in \operatorname{argmin}_{q\in\mathcal{I}_r} \mathrm{KL}\left(q; p_{Z|Y}(\cdot|\mathbf{y};\psi^*)\right) \text{ and } \psi^* \in \operatorname{argmax}_{\psi\in\Psi} F(q^*,\psi)\right\}. \tag{20}$$

We will prove that under **A1, A2(i)** and **A3**, $\mathcal{L}$ is the solution set of the generalized VEM algorithm and of the VEM algorithm (see Theorem 1 below; see also [4, Theorem 2], where the same result is obtained under different sufficient conditions). When $L$ is continuously differentiable on $\mathcal{I}_r \times \Psi$, the solution set of VEM (and of generalized VEM) can also be characterized as the set of the stationary points of $L$ in the interior of $\mathcal{I}_r \times \Psi$ (Appendix 8, Remark 6).

**A4** Assume either that

(i) the set $L(\mathcal{L})$ is compact.

(ii) for all compact $\mathcal{K} \subset \mathcal{I}_r \times \Psi$, $L(\mathcal{K} \cap \mathcal{L})$ is finite.

Under **A1-A2(i)**, $L$ is continuous on $\mathcal{I}_r \times \Psi$ and $\mathcal{L}$ is a closed subset of $\mathcal{I}_r \times \Psi$. Hence, **A4** is verified whenever $L(\mathcal{L})$ is bounded.

## 4.2 Monte-Carlo approximations

Under **A2(iii)** (resp.**A2(iv)**), MCVEM updates the $\beta$-component by computing the root of

$$\beta \mapsto \sum_{\mathbf{z}} \nabla_\beta H(\mathbf{z};\beta) q^{t+1}(\mathbf{z}) - \sum_{j=1}^{J_t} \nabla_\beta H(\mathbf{z}^j;\beta) \, \frac{\omega(\mathbf{z}^j;\beta,\beta^t)}{\sum_{r=1}^{J_t} \omega(\mathbf{z}^r;\beta,\beta^t)},$$

where $\omega(\mathbf{z};\beta,b) \propto \frac{p_Z(\mathbf{z};\beta)}{p_Z(\mathbf{z};b)}$. Hence the expectation $\nabla \ln W(\beta) = -\sum_{\mathbf{z}} \nabla_\beta H(\mathbf{z};\beta)\ p_Z(\mathbf{z};\beta)$ is estimated, in MCVEM, by importance reweighting the output $\{Z^{j,t}\}_j$ from the chain with stationary distribution $p_Z(\cdot;\beta^t)$. The update of the $\beta$-variable thus follows the MCMCML algorithm proposed by [12] for the estimation of fully observed Markov Random Field prior parameters.

We formulate sufficient conditions that imply a local uniform control of the difference between $\nabla \ln W$ and its Monte-Carlo approximation. Let $\mathbb{E}_{\lambda,\beta}$ be the expectation on the canonical space associated to the Markov random field with initial distribution $\lambda$ and stationary distribution $p_Z(\cdot;\beta)$. Let $\mathrm{Cl}(\mathcal{C}_\alpha)$ be the closure of the $\alpha$-neighborhood of some (bounded) set $\mathcal{C}$.

**A5** There exist $r \geq 2$ and a probability distribution $\lambda$ on $\mathcal{Z}$ such that for any compact subset $\mathcal{C} \subset \mathcal{B}$ and any $\alpha > 0$,

$$\sup_{\beta \in \mathrm{Cl}(\mathcal{C}_\alpha), b \in \mathcal{C}} \sup_{J \geq 1} J^{r/2} \mathbb{E}_{\lambda,b} \left[ \left| \nabla_\beta \{ \ln \tilde{W}^{J,b}(\beta) - \ln W(\beta) \} \right|^r \right]$$

is finite.

**A5** is verified whenever

$$\sup_{\beta \in \mathrm{Cl}(\mathcal{C}_\alpha), b \in \mathcal{C}} \sup_{J \geq 1} J^{r/2} \, \mathbb{E}_{\lambda,b} \left[ \left| \tilde{W}^{J,b}(\beta) - W(\beta) \} \right|^r \right]$$

is finite and

$$\sup_{\beta \in \mathrm{Cl}(\mathcal{C}_\alpha), b \in \mathcal{C}} \sup_{J \geq 1} J^{r/2} \, \mathbb{E}_{\lambda,b} \left[ \left| \nabla_\beta \{ \tilde{W}^{J,b}(\beta) - W(\beta) \} \right|^r \right]$$

is finite. Observe that both of these integrals are on the form

$$\mathbb{E}_{\lambda,b} \left[ \left| \sum_{j=1}^{J} \{ \mathcal{H}(Z^j; \beta, b) - \sum_{\mathbf{z}} \mathcal{H}(\mathbf{z}; \beta, b) p_Z(\mathbf{z}; b) \} \right|^r \right],$$

where $p_Z(\mathbf{z}; b)$ is the invariant probability distribution of the Markov chain $\{Z^j\}_j$ with initial distribution $\lambda$. Sufficient conditions implying this uniform control of the $L^r$-norm difference for a Markov chain can be found in [15] (see section 5 for an example). Finally, we assume that the number of simulations $\{J_t\}_t$ increases at a rate such that the larger $r$, the weaker the rate.

**A6** $\{J_t\}_t$ is a positive integer-valued sequence such that $\sum_{t \geq 0} J_t^{-r/2} < \infty$ where $r$ is given by **A5**.

We refer to section 5 for an illustration of a suitable choice of $\{J_t\}_t$ in practice.

## 4.3 A convergence theorem for some generalized VEM

Consider the generalized VEM algorithm that replaces, at each iteration, the global optimization in (16) by a local one on $\mathcal{B}^t$.

**Theorem 1** *Assume* **A1, A2(i)-A2(iii)** *and* **A3**. *Fix a positive sequence* $\{\gamma^t\}_t$ *such that* $\inf_t \gamma^t > 0$ *and let* $\{(q^t, \psi^t)\}_t$ *be the generalized VEM path started at* $(q^0, \psi^0) \in \mathcal{I}_r \times \Psi$. *The sequence* $\{L(q^t, \psi^t)\}_t$ *converges monotonically to* $L^* = L(q^*, \psi^*)$ *for some* $(q^*, \psi^*) \in \mathcal{L}$. *Furthermore, the sequence* $\{(q^t, \psi^t)\}_t$ *converges to the set* $\{(q, \psi) \in \mathcal{L}, L(q, \psi) = L^*\}$.

Our generalized VEM procedure applied with $\gamma^t = +\infty$ is the VEM procedure, so that Theorem 1 also addresses the convergence of VEM. Theorem 1 shows that the generalized VEM algorithm has the same asymptotic behavior as the VEM algorithm, since the limit set does not depend upon the sequence $\{\gamma^t\}_t$. In addition, this result coincides with [4, Theorem 2, assertion (1)] (the assertions (2) and (3) of this reference do not apply to VEM).

Our assumptions imply the conditions in [4]: under **A1, A2(i)-(iii)** and **A3**, the sequence is compact, this compact is stable under action of the point-to-set map associated to each iteration of VEM and this map is closed (see e.g. [4, Definition 1]) so the conditions of [4, Theorem 2] are verified. Our conditions, relative to the quantities defining the model, are more explicit than the conditions in [4] and easily checked in the considered applications (see Appendix 10).
Even though Theorem 1 can be a consequence of [4, Theorem 2, assertion (1)], we provide a detailed proof in Appendix 8. We indeed establish auxiliary results (closedness of the limit set $\mathcal{L}$, existence of Lyapunov functions, $\cdots$) that are crucial in the proof of Theorem 2 which is the original theoretical result of this contribution. These auxiliary results are not provided in [4].

## 4.4 An almost-sure convergence theorem for the stable MCVEM algorithm

The convergence of the random trajectories is established almost surely with respect to $\bar{\mathbb{P}}$, the probability on the canonical space associated to the trajectories of stable MCVEM, started at $(q^0, \psi^0)$, given the initial distribution $\lambda$ of the Markov random field, the sequence of compact sets $\{\mathcal{C}^t\}_t$ satisfying (17) and the sequence $\{\gamma^t\}_t$.

**Theorem 2** *Assume* **A1** *to* **A6**. *Let* $\{\mathcal{C}^t\}_t$ *be a sequence of compact sets satisfying (17),* $(q^0, \psi^0) \in \mathcal{I}_r \times \mathcal{C}^0$ *and* $\lambda$ *be given in* **A5**. *Fix a positive sequence* $\{\gamma^t\}_t$ *such that* $\inf_t \gamma^t > 0$. *Consider the stable MCVEM random sequence* $\{(q^t, \psi^t)\}_t$. *Then,*

*(i) (a)* $\lim_t \tau^t < \infty$ *w.p.1 and* $\limsup_t |\psi^t| < \infty$ *w.p.1.*

*(b)* $\{L(q^t, \psi^t)\}_t$ *converges w.p.1 to a connected component of* $L(\mathcal{L})$.

*(ii) If in addition* $L(\mathcal{L} \cap \mathrm{Cl}(\{(q^t, \psi^t)\}_t))$ *has an empty interior, then* $\{L(q^t, \psi^t)\}_t$ *converges w.p.1 to* $L^*$ *and* $\{(q^t, \psi^t)\}_t$ *converges to the set* $\mathcal{L}_{L^*} = \{(q, \psi) \in \mathcal{L}, L(q, \psi) = L^*\}$.

This theorem states that for almost all trajectories of stable MCVEM, the number of re-projections is finite and the path remains in a compact set. Furthermore, we deduce from Theorems 1 and 2 that when $L(\mathcal{L} \cap \mathrm{Cl}(\{(q^t, \psi^t)\}_t))$ has an empty interior, the stable MCVEM algorithm, the generalized VEM algorithm and the VEM algorithm have the same asymptotic behavior: in all cases, the sequence $\{L(q^t, \psi^t)\}_t$ converges to $L^* = L(q^*, \psi^*)$ for some $(q^*, \psi^*)$ in the solution set $\mathcal{L}$, and the path $\{(q^t, \psi^t)\}_t$ produced by the stable MCVEM path, the generalized VEM one or the VEM one, converges to some subset of $\mathcal{L}$. For example, if **A4(ii)** is verified, $L(\mathcal{L} \cap \mathrm{Cl}(\{(q^t, \psi^t)\}_t))$ is finite thus having an empty interior.

### 4.5 Extensions

*Finite state space $\mathcal{Z}$:* we assumed that the state space $\mathcal{Z}$ of the hidden MRF is finite for simplicity. Nevertheless, **A1** and **A2(i)** can be replaced by the conditions that KL and $L$ are continuous functions on $\mathcal{I}_r \times \Psi$ and the partition function $W$ is continuous on $\mathcal{B}$.
*Deterministic sequence $\{\gamma^t\}_t$ in Theorem 1:* the convergence theorem is stated for a given deterministic sequence $\{\gamma^t\}_t$ such that $\inf_t \gamma^t > 0$. This condition is crucial in the proof of Theorem 1 since it ensures a minimal growth of the Lyapunov function outside compact sets (see Appendix 8.2 and the use of $T^*$). Nevertheless, $\gamma^t$ could be chosen 'on line', as a function of the algorithm. For example, a natural choice is to determine $\gamma^t$ as a function of the Hessian $\nabla^2 \mathcal{Q}_2^{t+1}(\beta^t)$, provided that $\inf_t \gamma^t > 0$.
*Deterministic sequence $\{\gamma^t\}_t$ in Theorem 2:* here again, one could choose $\gamma^t$ as a function of $\nabla^2 \tilde{\mathcal{Q}}_2^{t+1}(\beta^t)$; in that case, $\{\gamma^t\}_t$ is a random sequence. The present proof of Theorem 2 is not valid anymore and conditions ensuring the sequence $\{\gamma^t\}_t$ is close in some sense to the sequence computed from the exact Hessian $\nabla^2 \mathcal{Q}_2^{t+1}(\beta^t)$ have to be assumed. Details of this extension are left to the interested reader.

## 5 Application to image segmentation

In this section, we turn more specifically to the applications. We consider simple models and use a $K$-color Potts model as the distribution of the hidden fields. Each $z_i$ takes one of $K$ states, which can represent $K$ different class assignments. Each of them is represented by a binary vector of length $K$ with one component being 1, all others being 0. The distribution of a $K$-color Potts model is defined by,

$$p_Z(\mathbf{z};\beta) \quad = \quad W(\beta)^{-1} \exp(\beta \sum_{i \sim j} z_i^t z_j) \ ,$$

where the notation $i \sim j$ represents all couples of sites $(i,j)$ which are neighbors. Parameter $\beta$ is a spatial parameter that controls the strength of the interaction between neighboring sites. In a segmentation framework, the Potts model acts as a regularizing (smoothing) term. The lower $\beta$, the weaker the regularization. The factorized conditional distribution $p_{Y|Z}(\mathbf{y} \mid \mathbf{z};\ \theta)$ is of the form $p_{Y|Z}(\mathbf{y} \mid \mathbf{z};\ \theta) = \prod_{i \in S} f_i(y_i \mid z_i;\ \theta)$ where $f_i$ is a univariate Gaussian distribution: if $z_i$ is in class $k$, $f_i$ is the Gaussian distribution with parameters $\mu_k$ and $\sigma_k$, $\mu_k$ and $\sigma_k$ being the mean and the standard deviation. The parameter to be estimated is then $(\beta, \theta)$ with $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$.
For the simulation step of *MCVEM*, we use the Gibbs sampler. For such models and this sampler, we show in Appendix 10 that the various assumptions are satisfied so that the previous convergence results applied : any MCVEM path converges, and the set of the limiting values is the set of the limiting values of VEM. Furthermore, assumption A5 is actually satisfied with any initial distribution $\lambda$ and any $r \geq 2$. Hence, there are no restrictions for the initialization of the Markov chains at each iteration, and suitable choices of $J_t$ are any polynomially increasing sequences.

We compare MCVEM to different algorithms when applied to parameter estimations and image segmentations. We first run an EM procedure (hereafter called *ind-EM*), assuming that missing data are independent, in order to illustrate the gain in taking into account the spatial information. The following other procedures are based on models assuming dependencies. As a typical simulation method, we run a kind of Monte-Carlo EM (hereafter

*MC2-EM*) where two Monte-Carlo approximations are introduced at each iteration. The first one corresponds to the MCEM algorithm [39] and the second one makes the M-step tractable by approximating the partition function $W(\beta)$ as in MCVEM. By combining the convergence results of MCVEM (Section 4) and of MCEM [15], it can be established that *MC2-EM* converges almost-surely to the stationary points of the incomplete log-likelihood $\ln p_Y(\mathbf{y};\psi)$ and due to its stochastic nature, converges to a (local) maximum [15]. *MC2-EM* has a much higher computational cost but it provides reference solutions to assess the proximity of the MCVEM limiting values to the maxima of the incomplete log-likelihood. We then compare to the *Mean Field* algorithm of [6] (see Section 3.1), as a typical deterministic variational algorithms. Finally, we run two other algorithms designed to overcome the intractability of EM in hidden MRF, the Gibbsian-EM [7] that combines Monte-Carlo techniques and pseudo-likelihood approximation and the *Simulated Field* of [6] (see Section 3.1). The latter two can also be seen as combinations of simulations and deterministic approximations but are not part of the novel strategy we propose. No convergence results are available for them.

In addition to parameter estimation, the way the segmentation task is carried out in the different procedures can varies. For *MCVEM*, *Mean Field* and *Simulated Field* algorithms, images are restored by using the MAP (Maximum a Posteriori) principle based on the factorized distribution $q^{t+1}$ that approximates the conditional distribution $p_{Z|Y}(\cdot|\mathbf{y};\psi^t)$. *Gibbsian-EM* and *MC2-EM* both generate realizations of the conditional field and the image reconstruction is performed using the MPM decision rule (Maximizer of the Posterior Marginal, [23]). Note that for the first three algorithms, the MAP and MPM rules coincide when applied to $q^{t+1}$ since $q^{t+1}$ is a factorized distribution.
For the Potts models, we assumed a first order neighborhood (four neighbors per pixel). For the stochastic algorithms (*i.e.* all but *ind-EM* and *Mean Field*), we report the mean values of the estimates along the random path, where the mean is over the iterations after the burn-in period. Regarding the segmentation results, the error rate (*i.e.* the proportion of misclassified pixels) corresponds to the mean error rate computed after the burn-in period.

## 5.1 Practical implementation of MCVEM

Prior to any performance comparison, we discuss implementation details of MCVEM such as the initialization of the Markov chain at each iteration, the choice of the simulation scheme $J_t$ and of the sequence $\{\gamma^t\}_t$. As an illustration, the algorithm is run on a $133 \times 142$ noise-corrupted 2-color image (Figure 10). We used Gaussian densities with class-dependent variances so that the true noise parameters are $(\mu_1, \sigma_1) = (51, 130)$ and $(\mu_2, \sigma_2) = (255, 300)$. In Figure 1, we plot $(\mu_1^t, \sigma_1^t, \beta^t)$ as a function of the number of iterations when the Markov chain is, at each iteration, initialized at the same point (solid line) or at the last sample drawn at the previous iteration (dot line). The considered $J_t$ is $J_t \sim (2t)^{1.2}$. In the two cases, the results, in terms of parameter estimation and segmentation, are similar but the convergence when using the first strategy is very slow. The same observation holds for other choices of $J_t$ so that in what follows, only the second strategy will be kept.
We then consider different schemes for $J_t$, namely $J_t \sim (2t)^{1.01}$, $J_t \sim (2t)^{1.3}$ and $J_t \sim (2t)^{1.5}$. All schemes result in a convergence to the same value of $\beta$. A zoom on the evolution of the successive $\beta$ values (Figure 2) shows that the value of $\beta$ in average is not sensitive to the scheme but that its variation is all the smaller as the rate is higher (a phenomenon already mentioned in [15]). The mean values, computed when the curves stabilize, are equal to 0.93 while the standard deviations are respectively 1.7 $10e{-}3$, 0.9 $10e{-}3$ and 0.6 $10e{-}3$. Similar behavior was observed on other images suggesting not surprisingly that limiting the number of simulations has a cost in that it produces paths with higher variations. In the following developments, we will consider $J_t \sim (2t)^{1.3}$.

We then study the robustness of *MCVEM* to the choice of the starting parameter values and to the choice of $\gamma$. We consider in turn three sets of parameter starting values. The means and variances are first set to the empirical means and variances corresponding to a kmeans classification (displayed in Figure 10) and then to those corresponding to a 2-color classification obtained by simple thresholding of the image pixels values. For the first case, two values of $\beta$ are considered, $\beta = 1$ and $\beta = 5$ while for the second case, only $\beta = 5$ is used. For MCVEM, $\gamma$ is constant over the iterations, $\gamma^t = \gamma$ and is respectively set to $5\ 10^{-2}$ and $5\ 10^{-3}$ for $\beta = 5$ and $\beta = 1$. The path $\{\beta^t\}_t$ of successive estimations of $\beta$ is plotted in Figure 3. We observe that the estimation of $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$ is well performed whatever the algorithm. The plots show that the limiting behavior of *MCVEM* (dash-doted lines) does not depend on $\gamma$, at least when $\gamma$ is small enough. For large values of $\gamma$ (say $\gamma = 0.1$), the sequence $\{\beta^t\}_t$ may oscillate for a long time between two values of the form $\beta$ and $\beta+\gamma$. This illustrates the fact that $\tilde{W}^{J_t,\beta^t}$ can be considered as a reasonable approximation of $W(\beta)$ in a neighborhood of $\beta^t$, and justifies the introduction of a local optimization domain $\mathcal{B}^t$ in the update of $\beta$. This local optimization explains the linear path of *MCVEM* in the first iterations. These plots illustrate that *MCVEM* is very robust to initialization.
For comparison with the two other variational methods we consider, we also run the *Mean Field* and *Simulated Field* algorithms and show the results on the same figure 3. It appears that the starting value is crucial for the limiting behavior of *Mean Field*. On some other synthetic images (not shown here), *Mean Field* actually fails to converge even with reasonable initializations such as those provided by running a *k-means* algorithm. The trajectories of *Simulated Field* do not converge to some fixed limiting value but the behavior of the different trajectories is similar. The same kind of phenomenon was already pointed out in [2] for the *Restoration-Maximization* algorithm close in spirit to the *Simulated Field* algorithm. We believe that convergence of the *Simulated Field* algorithm has to be understood in a different way. An approach similar to what is done for the so-called Stochastic EM algorithm is more appropriate (see [28, 5]). The sequence $\{(q^t, \psi^t)\}_t$ is a realization of a Markov Chain and the asymptotic behavior of this sequence is related to the ergodic behavior of this Markov chain. Hence, averages of the parameters should converge and this suggests to replace the current implementation of *Simulated Field* algorithm by an averaging procedure [29]. However, such extensions are beyond the scope of this paper and we run the algorithm as described in [6]. Despite the variations in the estimation of the spatial parameter $\beta$, the corresponding segmentations are quite stable: the mean error rate is in the range $(2.86\%, 2.92\%)$ for *MCVEM* (resp. $(2.82\%, 3.10\%)$ for *Mean Field* and $(3.42\%, 3.65\%)$ for *Simulated Field*).

We finally discuss how the possible non-unicity of $q^{t+1}$, the mean-field approximation of the conditional field $p_{Z|Y}(\cdot|\mathbf{y}; \psi^t)$, may affect the resulting image segmentations. To that goal, we compute the mean error rates for the segmented images when $\beta$ is assumed to be known but $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$ is unknown. On Figure 4, we plot these mean error rates versus $\beta$ for two different starting points corresponding respectively to a kmeans and a thresholding classification as above. These plots show that for large values of $\beta$, the segmentation is greatly dependent of the initial segmentation. In addition, the curves give an idea of the $\beta$ value that corresponds to the minimum error rate. For *MCVEM* and *Simulated Field*, this naive computation is not far from the estimates obtained by running the full algorithms when all the parameters $(\beta, \theta)$ are unknown (these estimations are reported in Table 4). *MCVEM* converges to a limiting value and *Simulated Field* fluctuates around a mean value such that the segmentation is not affected by the non-unicity of $q^{t+1}$. This is not the case for *Mean Field* thus showing that the *Mean Field* segmentation may depend on the implementation of the algorithm.

## 5.2 Synthetic and real images

We now compare into more details the algorithms performance when applied to parameter estimation and image segmentation. We report the estimations of $\beta$ and $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$ and the mean segmentation error rates when a ground truth is available. For comparison, we also indicate in column *ref.* the error rates when the parameters are not estimated but fixed to their true values if known. When given, the corresponding segmentations are computed after the same fixed number of iterations (200) for each iterative algorithm. Three types of test images are presented. Detailed comments on the results are postponed after the description of all experiments.

The algorithms are first tested on images simulated from hidden Potts models for which the true parameters $\beta$ and $\theta$ are known. We created $100 \times 100$ images by simulating 2D $K$-color Potts models for $K = 2, 3, 4$ and different values of $\beta$ (lower than the critical value $\beta_c = \ln(1 + \sqrt{K})$), and then adding a Gaussian noise. For each set of parameters we investigate, 20 realizations of each corresponding Potts model are simulated. We then run the different algorithms on these 20 simulations. The results are reported in Tables 1, 2 and 3. For $K = 2$, the values reported are the mean and standard deviation values over the 20 runs. For $K = 3, 4$, we observe that the estimation of the parameter $\theta$ is always satisfying and only the results on $\beta$ are reported.

The following test images are noise-corrupted images corresponding to known values of $K$. These images before degradation are not realizations from a known Markov field model. The first image is the logo image described in 5.1 and shown in Figure 10. The other example is a $128 \times 128$ image obtained by adding some Gaussian noise to the 4-color top left image of Figure 10. The noise parameters are given by $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, 4\}$ with $\mu_k = k$ and $\sigma_k = 0.5$ for $k = 1, \ldots, 4$. The results are reported in Tables 4 and 5. The corresponding segmentations are shown in Figures 10 and 10.

We finally run the algorithms on real images for which a true value of $K$ does not exist (in real-life, it is usually part of the problem to assess its value) but for which intuition or expert knowledge could give an indication of what would be a reasonable value. As an illustration, the top left image in Figure 7 is a $256 \times 256$ image of muscle fibers, the top left image in Figure 8 is a $76 \times 91$ PET image of a dog lung (see [34] for more details on its nature and origin) and the top left image in Figure 10 is a $256 \times 256$ satellite image. They have been chosen because they correspond to rather different application domains and because non expert users can easily assess the quality of their segmentations.
More specifically, a simple histogram of the grey-levels in the muscle image shows essentially three modes. The image contains dark color fibers which are homogeneous and more textured lighter color fibers. The region between the fibers is also light color but homogeneous. We chose this image to compare the segmentation results because it is relatively easy to give an ideal segmentation for this image. We run the algorithms for $K = 4$ because the lighter color fibers can clearly be divided in two groups in terms of their grey-levels. We then assess the algorithms ability to recover the two-three types of fibers against the background. The resulting segmentations are shown in Figure 7.
For the dog lung image, the aim is to distinguish the lung from the rest of the image in order to measure the heterogeneity of the tissue in the region of interest. Only pixels in this delimited area are then considered to compute a heterogeneity measure, such as a coefficient of variation. The interpretation of the image suggests that 3-color segmentations are reasonable. The image is constructed based on radioactive emissions from gas in the lung. Ideally, the background should correspond to one color and two other colors should account for the high gas density in the interior of the lung and the somewhat lower gas

density around the periphery. The resulting segmentations are shown in Figure 8.
Figure 10 is a SPOT satellite image representing part of the Aquitaine region in France. It contains large homogeneous regions (large fields, woods ),precise contours (rivers, roads) and more heterogeneous areas (houses, small fields) or textured parts. Whether relevant segmentations should focus on contours or regions may depend on the application in mind.

All tables show that *ind-EM* differs from the other algorithms: the estimates are somewhat poor (see *e.g.* Table 5) and the error rates are much higher. The gain in taking into account spatial dependencies clearly appears.

We observe that the estimation of the means and standard deviations $\{(\mu_k, \sigma_k), k = 1, \dots K\}$ is an easy task in the sense that all algorithms (except *ind-EM*) have similar good performances. We then focus our comments on the estimation of the spatial parameter $\beta$ which is more critical. When the true value is lower than the critical value $\beta_c$, *MCVEM* seems to underestimate $\beta$ (see Tables 1-3). More generally, *MCVEM* provides the lowest estimates, while *Mean Field* provides the highest ones. The *Mean Field* algorithm systematically overestimates $\beta$. It is quite difficult to determine which approach is the best, since the value of the spatial parameter $\beta$ acts upon the image segmentation. Nevertheless, the results of *MC2-EM* which converges to the (local) maxima of the incomplete log-likelihood $\ln p_Y(\mathbf{y}; \psi)$ can be taken as reference values. It appears that *MCVEM* and *MC2-EM* are very close (see Tables 5 and 4) while *Mean Field* and *Simulated Field* and *Gibbsian-EM* are of a different kind. For the $\beta$ estimation, *Simulated Field* is close to *Gibbsian-EM* while *Mean Field* is the most atypical. Despite *Simulated Field* and *Gibbsian-EM* rely on different tools (mean-field based variational technique on one hand, pseudo-likelihood approximation on the other hand), they are numerically close. We believe that, due to the ergodicity of the discrete-valued Markov chain which admits the conditional field $p_{Z|Y}$ as invariant distribution, they have indeed very similar asymptotic behaviors.

In terms of segmentation results, *MCVEM* leads to very satisfying error rates: for the hidden Potts images, the error rates are close to the minimal error rates (achieved with *MC2-EM* and *Gibbsian-EM*) even though the $\beta$ estimate is poorer. The algorithms divide into two groups: on one hand, *MCVEM* and *MC2-EM* which provide lower values of $\beta$ and consequently images with possibly more isolated points (Figures 10 and 10); on the other hand the *Mean Field*, *Simulated Field* and *Gibbsian-EM* algorithms that provide larger $\beta$ estimates and smoother images. It appears clearly, *e.g.* on the logo image, that *MCVEM* tends to better preserve fine structures, the continuous lines in the original image being less interrupted in various locations (see also the satellite image). It performs slightly better than *Simulated Field* and *Mean Field*. The triangle image with no such fine structures cannot illustrate this ability of the algorithm. However we observed the same phenomenon on various other synthetic images with fine structures. On the contrary, when large homogeneous area exists, *MCVEM* and *MC2-EM* segmentations are not smooth enough and isolated points are still visible, producing consequently slightly higher error rates (Figure 10 and Table 5). Note that in practice such points are not an issue since they can be easily dealt with afterwards using some simple morphological operator leading to potentially further improved error rates. For example, application of a median filter on the *MCVEM* image reconstruction improves the error rate, 2.73% instead of 2.89% for the logo image (Figure 10, bottom right), 0.63% instead of 0.81% for the triangle image (Figure 10, bottom right). Similar conclusions can be drawn from the real image experiments. The *MCVEM*, *Simulated Field*, *MC2-EM* and *Gibbsian-EM* algorithms perform similarly well while the *Mean Field* algorithm has trouble segmenting some of the light color fibers correctly (Figure 7). For the dog lung image (Figure 8), the *MCVEM* and *MC2-EM* segmentations are not as smooth when considering the light grey region but provide a more accurate segmentation of the white region. For instance, the segmentation of the upper and central parts of the right lung

looks better. All spatial algorithms provide however smoother segmentations than *kmeans* and *ind-EM*. For the *Simulated Field* algorithm, we report the segmentations corresponding to the implementation of [6] as specified in Section 3.1; however, for these two images, 200 iterations are not enough for convergence. We observe that when carrying out more simulations at each iteration or similarly when performing more iterations, the *Simulated Field* algorithm tends to loose the small regions (*e.g.* the central and background small regions in Figure 8) so that the segmentations are then very close to the *Gibbsian-EM* ones.

As mentioned earlier, for each algorithm, the displayed segmentation is computed using the current state of the algorithm after a fixed number of iterations. The error rate computed at each iteration, after the burn-in period, stabilizes for *Mean Field* and *MCVEM*. For example, on the logo image of Figure 10, the error rate is almost always constant. For fixed values of the parameters, the *MCVEM* segmentation procedure does not require simulations any more and is equivalent to the *Mean Field* segmentation procedure. This is not true for the *MC2-EM*, *Gibbsian-EM* and the *Simulated Field* procedures which remain stochastic since even for fixed values of the parameters, the segmentation step still relies on samples drawn from the conditional field. Nevertheless, for *MC2-EM* and *Gibbsian-EM*, the error rate has a small variation along the path (4 $10e{-}2$ for the logo image) while the *Simulated Field* algorithm provides the most unstable procedure since, as already mentioned, its paths do not converge. For the logo image, the error rate variation is 11 $10e{-}2$. More complex segmentation rules could be considered to overcome this instability. For example, the different segmentations that can be computed along the iterations can be seen as successive votes, and the final image reconstruction based on the mean value of these votes. For the logo image, this yields for resp. the *Simulated Field*, the *Gibbsian-EM* and the *MC2-EM* algorithms a mean error rate of 3.18%, 2.94%, 2.83% and a lower variation along the path (resp. 2.40 $10e{-}2$, 1.70 $10e{-}2$ and 1.25 $10e{-}2$).

# 6 Discussion and future work

In this paper, we proposed a new algorithm to carry out Markov model-based segmentation in practice, combining variational and MCMC ideas. This combination allowed us to prove the first, to our best knowledge, convergence result for this kind of algorithms. This result extends to a whole new class of algorithms. It is based on the idea of seeing the algorithm under study as a perturbed version of a reference algorithm for which convergence results are well established and usually based on a well identified Lyapunov function. For instance, this applies when the model complexity leads to an exact deterministic algorithm which is intractable and must be replaced for practical implementation by an approximate version. The key idea in our contribution, is that although a Lyapunov function does not usually exist for the perturbed algorithm, it is possible to control the distance to this Lyapunov function. Studying its limit set is then made possible through the definition of a set such as $\mathcal{L}$ in Section 4.1, which defines the algorithm solutions as satisfying an optimality criterion. These observations open the way to a general approach to implement intractable (deterministic) algorithms in practice through adequately designed stochastically perturbed versions. In the hidden Markov random fields context, a natural development of the present work would then be to further study other noisy EM versions with preserved limit sets.

As regards the *MCVEM* algorithm we focused on, we showed that in addition to garanteed convergence properties, it provided good segmentation results and compared favorably to other approximated algorithms. Various experiments pointed out that *MCVEM* was close

to the *MC2-EM* algorithm based on the *MCEM* algorithm which is known to converge to local maxima of the incomplete log-likelihood. *MCVEM* is then clearly to be favored since it has a much lower computational cost than *MC2-EM*. In particular, the segmentation step in *MCVEM* is simple and does not require the additional computations needed in *MC2-EM*. Also *MCVEM* tends to provide adequate regularizations through values of $\beta$ which are not too large and has this way the ability to preserve fine structures. This characteristic can also be responsible for misclassified pixels but they mainly correspond to isolated points. These points can be easily dealt with using some straightforward postprocessing procedure. The performance of *MCVEM* is then very satisfying, all the more so as the results could be further improved by more focus on the use of better sampling techniques. For illustration purpose, we restricted to a simple Gibbs sampler but investigating the use of more sophisticated methods (*e.g.* [33, 16]) would be worthwile. More generally, an alternative approach of the sampling problem would be to consider stochastic approximation techniques such as presented and used in [42] and [10]. We suspect the same kind of convergence results could follow using the same idea of controlling the distance to a reference Lyapunov function.

In this paper, comparison with other existing EM-like procedures showed that the relationship between our algorithm and the former was not obvious. Our study revealed three groups. *MC2-EM* and *MCVEM* distinguish from the *Gibbsian-EM* of [7] and from the *Mean Field* and *Simulated Field* algorithms of [6]. *Simulated Field* does not converge in the same sense and is closer to the *Gibbsian-EM*. It tends to produce smoother segmentations but more unstable trajectories. *Mean Field* has a third specific behavior. Its convergence is not always guaranteed and when observed, the resulting segmentations are very smooth. Further comparisons and investigations would be useful. We believe this first effective step opens the way to a better understanding of the behavior and theoretical properties of a lot of Markov model based algorithms. In particular, analysing how simulation steps should be incorporated so as to interact advantageously with deterministic approximations seems promising.

# 7 Acknowledgments

The authors would like to thank Juliette Blanchet for help with some experiments, Marc Sigelle and Olivier Cappé for fruitful discussions.

# 8 Proof of Theorem 1

A map $T$ from points of $\mathcal{I}_r \times \Psi$ to subsets of $\mathcal{I}_r \times \Psi$ is called a point-to-set map on $\mathcal{I}_r \times \Psi$. Let $T^t$ be the point-to-set map on $\mathcal{I}_r \times \Psi$ in the $(t+1)$-th generalized VEM iteration: $(\bar{q}, \bar{\psi}) \in T^t(\tilde{q}, \tilde{\psi})$ where $\tilde{\psi} = (\tilde{\theta}, \tilde{\beta})$ iff

$$\bar{q} \in \operatorname{argmin}_{q \in \mathcal{I}_r} \mathrm{KL}(q; p_{Z|Y}(\cdot|\mathbf{y}; \tilde{\psi})),$$
$$\bar{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \sum_{\mathbf{z} \in \mathcal{Z}} \ln p_{Y|Z}(\mathbf{y}|\mathbf{z}; \theta)\ \bar{q}(\mathbf{z}),$$
$$\bar{\beta} = \operatorname{argmax}_{b \in \{x, |x - \tilde{\beta}| \leq \gamma^t\}} \sum_{\mathbf{z} \in \mathcal{Z}} \ln p_Z(\mathbf{z}; b)\ \bar{q}(\mathbf{z}).$$

Let $T^*$ be defined as $T^t$ with $\gamma^t$ replaced by $\inf_t \gamma^t$ in the update of the $\beta$-component. Under A1 and A2(i)-(iii), $T^t$ and $T^*$ are well-defined.

## 8.1 Auxiliary results

**Definition 3** *Let $\{T^t\}_t$ and $T^*$ be point-to-set maps on $\mathcal{I}_r \times \Psi$ and let $\mathcal{L} \subset \mathcal{I}_r \times \Psi$. $L : \mathcal{I}_r \times \Psi \to \mathbb{R}_+$ is a Lyapunov function relative to $(\{T^t\}_t, \mathcal{L})$ iff for all $t$, (a) for any $u \in \mathcal{I}_r \times \Psi$, $v \in T^t(u)$, $L(v) \geq L(u)$; (b) for any compact set $\mathcal{K} \subset (\mathcal{I}_r \times \Psi) \setminus \mathcal{L}$, $\inf_{u \in \mathcal{K}, v \in T^t(u)} \{L(v) - L(u)\} > 0$.*

This definition is more restrictive than the definition given for example in [44] and [4]: in these contributions, the condition $(b)$ is substituted by the condition $\{L(v) - L(u)\} > 0$ for all $(u, v)$ such that $u \notin \mathcal{L}$ and $v \in T^t(u)$.

**Lemma 4** *Assume A1, A2(i)-(iii) and A3. The function $L$ is a positive continuous Lyapunov function relative to $(\{T^t\}_t, \mathcal{L})$ and to $(T^*, \mathcal{L})$.*

*Proof:* Under A1 and A2(i), $L$ is continuous on $\mathcal{I}_r \times \Psi$. From (8) and the definition of $q^{t+1}$, $F(q^{t+1}, \psi^t) \geq F(q^t, \psi^t)$. By definition of $\psi^{t+1}$, $F(q^{t+1}, \psi^{t+1}) \geq F(q^{t+1}, \psi^t)$. Hence, for any $t$, $F(q^{t+1}, \psi^{t+1}) \geq F(q^t, \psi^t)$. If $(q^t, \psi^t) \notin \mathcal{L}$, the inequality gets strict. We now prove that this inequality remains strict, uniformly when $u \in \mathcal{K}$. Since $L$ is continuous, it is sufficient to prove that $T^t(\mathcal{K})$ is in a compact subset of $\mathcal{I}_r \times \Psi$. If $(\bar{q}, \bar{\psi}) \in T^t(\mathcal{K})$, $(\bar{q}, \bar{\psi}) \in \{(q, \psi) \in \mathcal{I}_r \times \Psi, L(q, \psi) \geq \inf_{\mathcal{K}} L\}$ which is a compact set of $\mathcal{I}_r \times \Psi$ by A3. This concludes the first part of the proof. The second part, relative to $T^*$ is along the same lines and is omitted. □

**Lemma 5** *Under A1 and A2(i)-(iii), $\mathcal{L}$ is a closed subset of $\mathcal{I}_r \times \Psi$.*

*Proof:* The functions $(q, \psi) \mapsto \mathrm{KL}(q; p_{Z|Y}(\cdot|\mathbf{y}; \psi))$ and $(q, \psi) \mapsto F(q, \psi)$ are continuous on $\mathcal{I}_r \times \Psi$. It is thus trivial to verify that any point $(q, \psi) \in \mathcal{I}_r \times \Psi$ which is the limit point of a converging $\mathcal{L}$-valued sequence, is in $\mathcal{L}$. □

**Remark 6** *Under additional regularity conditions on the model, Lemmas 4 and 5 can be proved with $\mathcal{L}$ replaced by the set of the stationary points of $L$ in the interior of $\mathcal{I}_r \times \Psi$.*

## 8.2 Conclusion

Under A3, $\mathcal{K} = \{(q, \psi) \in \mathcal{I}_r \times \Psi, L(q, \psi) \geq L(q^0, \psi^0)\}$ is a compact subset of $\mathcal{I}_r \times \Psi$; since $L(q^{t+1}, \psi^{t+1}) \geq L(q^t, \psi^t) \geq \cdots \geq L(q^0, \psi^0)$, $\{(q^t, \psi^t)\}_t$ is in $\mathcal{K}$ and the generalized VEM path is compact.

The first claim of Theorem 1 is that the sequence $\{L(q^t,\psi^t)\}_t$ converges monotonically to $L^*$. This comes from the existence of a continuous Lyapunov function and can be proved along the same lines as the proof of [44, Lemma 4.1. p.89]. The details are omitted.
We now establish the existence of some $(q^*,\psi^*) \in \mathcal{L}$ such that $L^* = L(q^*,\psi^*)$. Define the set $\mathcal{A} = L(\mathcal{L} \cap \mathcal{K})$, which is compact in $\mathbb{R}_+$ since $L$ is continuous on $\mathcal{I}_r \times \Psi$, $\mathcal{K}$ is compact in $\mathcal{I}_r \times \Psi$ and $\mathcal{L}$ is closed in $\mathcal{I}_r \times \Psi$ (Lemma 5). Let $\alpha > 0$ and set $\mathcal{A}_\alpha$ be the $\alpha$-neighborhood of the closed set $\mathcal{A}$ in $L(\mathcal{I}_r \times \Psi)$. As $\mathcal{A}$ is compact, $\mathcal{A} = \bigcap_{\alpha>0} \mathcal{A}_\alpha$. Since $\mathcal{A}_\alpha$ is a finite union of disjoint bounded open intervals, there exist $n_\alpha \geq 0$ and two increasing real valued sequences $\{a_\alpha(k)\}$ and $\{b_\alpha(k)\}$, $1 \leq k \leq n_\alpha$, such that

$$\mathcal{A}_\alpha = \bigcup_{k \in \{1,\cdots,n_\alpha\}} (a_\alpha(k), b_\alpha(k)). \tag{21}$$

$L^{-1}(\mathcal{A}_\alpha)$ is an open neighborhood of $\mathcal{L} \cap \mathcal{K}$, and we define

$$\epsilon_\alpha = \inf_{u \in \mathcal{K} \setminus L^{-1}(\mathcal{A}_\alpha), v \in T^*(u)} \{L(v) - L(u)\}. \tag{22}$$

Since $\mathcal{K} \setminus L^{-1}(\mathcal{A}_\alpha)$ is a compact subset of $\mathcal{I}_r \times \Psi$, $\epsilon_\alpha$ is positive by Lemma 4. By definition of $T^*$, $L(q^{t+1},\psi^{t+1}) \geq L(\bar{q}^{t+1},\bar{\psi}^{t+1})$ where $(\bar{q}^{t+1},\bar{\psi}^{t+1}) \in T^*(q^t,\psi^t)$; together with (22), this implies

$$(q^t,\psi^t) \in \mathcal{K} \setminus L^{-1}(\mathcal{A}_\alpha) \Longrightarrow L(q^{t+1},\psi^{t+1}) - L(q^t,\psi^t) \geq \epsilon_\alpha. \tag{23}$$

Define $k^\star_\alpha = \min\{1 \leq k \leq n_\alpha, \limsup_t\ L(q^t,\psi^t) < b_\alpha(k)\}$ and $I(\alpha) = (a_\alpha(k^\star_\alpha); b_\alpha(k^\star_\alpha))$. Since $\{L(q^t,\psi^t)\}_t$ is bounded, (23) shows that $\{L(q^t,\psi^t)\}_t$ is infinitely often (i.o.) in $\mathcal{A}_\alpha$, and since $\mathcal{A}_\alpha$ is a finite union of intervals, $\{L(q^t,\psi^t)\}_t$ is i.o. in an interval of (21); thus, $\limsup_t\ L(q^t,\psi^t) = \lim_t L(q^t,\psi^t) = L^* \in I(\alpha)$. Let $0 < \alpha_1 < \alpha_2$. By definition, $\mathcal{A}_{\alpha_1} \subset \mathcal{A}_{\alpha_2}$, thus $I(\alpha_1) \subset I(\alpha_2)$ and $L^* \in I(\alpha_1) \cap I(\alpha_2)$. Let $\{\alpha_n\}_n$ be a decreasing sequence such that $\lim_n \alpha_n = 0$; then $L^* \in \bigcap_n I(\alpha_n)$. $\{I(\alpha_n)\}_n$ is a decreasing sequence of intervals, $\bigcap_n I(\alpha_n)$ is an interval and $\bigcap_n I(\alpha_n) \subset L(\mathcal{L} \cap \mathcal{K})$. Hence, $\{L(q^t,\psi^t)\}_t$ converges to this interval, thus proving that $L^* = L(q^*,\psi^*)$ for some $(q^*,\psi^*) \in \mathcal{L}$.
Finally, the convergence of $\{(q^t,\psi^t)\}_t$ to a subset of $\mathcal{L}$ is a consequence of (23).

# 9 Proof of Theorem 2

The proof of Theorem 2 mimics the proof of [15, Theorem 3]. To that goal, we adapt the deterministic results by [15, Section 5.1] in order to take into account the fact that in the present contribution, **(a)** VEM map and MCVEM map are point-to-set maps, and **(b)** we have a family of "exact" map $\{T^t\}_t$. We start with stating these modified deterministic results (Appendix 9.1) and prove Theorem 2 in Appendix 9.2.

## 9.1 Deterministic results

The following three propositions are respectively adapted from [15, Propositions 9,10,11]. The first proposition provides sufficient conditions for the convergence of some perturbed iterative maps on $\mathcal{U}$, which approximate in some sense a sequence of maps having a Lyapunov function. The $\mathcal{U}$-valued path of the "perturbed" algorithm is assumed to be compact. The second proposition proves that this compactness assumption can be replaced by a recurrence condition whenever there exists a Lyapunov function that controls excursion outside

the compact sets of $\mathcal{U}$. As a corollary, the third proposition shows that the stabilization procedure derived in Section 3.2 ensures the compactness of the sequence defined by the perturbed maps.

**Proposition 7** *Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$, $\mathcal{K}$ be a compact subset of $\mathcal{U}$ and $\mathcal{L} \subseteq \mathcal{U}$ such that $\mathcal{L} \cap \mathcal{K}$ is compact in $\mathcal{U}$. Let $\{T^t\}_t$ and $T^*$ be point-to-set maps on $\mathcal{U}$. Let $L$ be a continuous Lyapunov function relatively to $(\{T^t\}_t, \mathcal{L})$ and to $(T^*, \mathcal{L})$ such that for all $t \geq 0$ and $u \in \mathcal{U}$, there exist $v^{t+1} \in T^t(u)$ and $w \in T^*(u)$ and $L(v^{t+1}) \geq L(w)$. Assume that there exists a $\mathcal{K}$-valued sequence $\{u^t\}_t$ such that there exists $v^{t+1} \in T^t(u^t)$ and $\lim_t |L(u^{t+1}) - L(v^{t+1})| = 0$. Then $\{L(u^t)\}_t$ converges to a connected component of $L(\mathcal{L} \cap \mathcal{K})$. If $L(\mathcal{L} \cap \mathcal{K})$ has an empty interior, $\{L(u^t)\}_t$ converges to $L^\star$ and $\{u^t\}_t$ converges to the set $\mathcal{L}_{L^\star} \cap \mathcal{K}$ where $\mathcal{L}_{L^\star} = \{u \in \mathcal{L}, L(u) = L^\star\}$.*

**Proposition 8** *Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$, $\{T^t\}_t$ and $T^*$ be point-to-set maps on $\mathcal{U}$ and $\mathcal{L} \subset \mathcal{U}$. Assume that* **(S1)** *there exists a continuous Lyapunov function $L$ relative to $(\{T^t\}_t, \mathcal{L})$ and to $(T^*, \mathcal{L})$ such that* ***(a)*** *for all $M > 0$, the level set $\{u \in \mathcal{U}, L(u) \geq M\}$ is compact in $\mathcal{U}$,* ***(b)*** *$\mathcal{U} = \bigcup_{n \geq \inf_{\mathcal{U}} L} \{u \in \mathcal{U}, L(u) \geq n\}$,* ***(c)*** *for all $u \in \mathcal{U}$, there exist $v^{t+1} \in T^t(u)$ and $w \in T^*(u)$ such that $L(v^{t+1}) \geq L(w)$.* **(S2)** *$L(\mathcal{L})$ is compact, or* **S2'** *$L(\mathcal{L} \cap \mathcal{K})$ is finite for all compact set $\mathcal{K} \subseteq \mathcal{U}$.* **(S3)** *there exists a $\mathcal{U}$-valued sequence $\{u^t\}_t$ such that* ***(a)*** *$\{u^t\}_t$ is infinitely often in a compact subset $\mathcal{C}^0 \subseteq \mathcal{U}$ and* ***(b)*** *for any compact set $\mathcal{K} \subseteq \mathcal{U}$, there exists $v^{t+1} \in T^t(u^t)$ such that $\lim_t |L(u^{t+1}) - L(v^{t+1})| \mathbb{1}_{u^t \in \mathcal{K}} = 0$.*
*Then $\{u^t\}_t$ is in a compact subset of $\mathcal{U}$.*

Let $\{\tilde{T}^t\}_t$ be a family of point-to-set maps on $\mathcal{U}$. Choose a sequence of compact subsets $\{\mathcal{C}^t\}_t$ of $\mathcal{U}$ such that for any $t \geq 0$, $\mathcal{C}^t \subsetneq \mathcal{C}^{t+1}$ and $\mathcal{U} = \bigcup_{t \geq 0} \mathcal{C}^t$. Define a sequence $\{u^t\}_t$ as follows: let $u^0 \in \mathcal{C}^0$ and set $\tau^0 = 0$.

- if $\tilde{T}^t(u^t) \subset \mathcal{C}^{\tau^t}$, choose $u^{t+1} \in \tilde{T}^t(u^t)$ and set $\tau^{t+1} = \tau^t$.
- else, set $u^{t+1} = u^0$ and $\tau^{t+1} = \tau^t + 1$.

**Proposition 9** *Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$, $\{T^t\}_t$ and $\{\tilde{T}^t\}_t$ be point-to-set maps on $\mathcal{U}$. Let $\{u^t\}_t$ be the sequence given by the re-projection procedure above. Assume* ***(a)*** *S1-2 holds,* ***(b)*** *for all $u \in \mathcal{C}^0$, for all $\tilde{u}^{t+1} \in \tilde{T}^t(u)$, there exists $v^{t+1} \in T^t(u)$ such that $\lim_t |L(\tilde{u}^{t+1}) - L(v^{t+1})| = 0$ and* ***(c)*** *for any compact subset $\mathcal{K} \subseteq \mathcal{U}$, for all $\tilde{u}^{t+1} \in \tilde{T}^t(u^t)$, there exists $v^{t+1} \in T^t(u^t)$ and $\lim_t |L(\tilde{u}^{t+1}) - L(v^{t+1})| \mathbb{1}_{u^t \in \mathcal{K}} = 0$. Then, $\limsup_t \tau^t < \infty$ and $\{u^t\}_t$ is a compact sequence.*

*Indications for the proof:* The proofs can easily be adapted from the proofs of [15, Propositions 9,10,11], up to a modification that plays a central role and we now detail.

In [15], the key assumption is the existence of a Lyapunov function $L$ relative to $(T, \mathcal{L})$, for some point-to-point map $T$. This implies that for any compact $\mathcal{K} \subset \mathcal{U}$, $\inf_{\mathcal{K}}\{L(T(u)) - L(u)\} \geq \epsilon > 0$, and yields an inequality which is fundamental in the proof: for any $u^t \in \mathcal{K}$,

$$L(u^{t+1}) - L(u^t) \geq L(u^{t+1}) - L(T(u^t)) + \epsilon.$$

In our case, we assume the existence of a Lyapunov function $L$ relative to $(\{T^t\}_t, \mathcal{L})$ and to $(T^*, \mathcal{L})$, for the point-to-set maps $\{T^t\}_t$ and $T^*$ such that for all $t \geq 0$ and $u \in \mathcal{U}$, there exist $v^{t+1} \in T^t(u)$ and $w \in T^*(u)$ and $L(v^{t+1}) \geq L(w)$. This implies that for any compact $\mathcal{K} \subset \mathcal{U}$, $\inf_{u\in\mathcal{K}, w\in T^*(u)}\{L(w) - L(u)\} \geq \epsilon > 0$, and for any $u^t \in \mathcal{K}$, there exists $v^{t+1} \in T^t(u^t)$ such that

$$L(u^{t+1}) - L(u^t) \geq L(u^{t+1}) - L(v^{t+1}) + \epsilon.$$

## 9.2 Conclusion

We prove the assertion (i)(a) and to that goal, we check the conditions of Proposition 9. $T^t$ refers to the $(t+1)$-th generalized VEM iteration and $\tilde{T}^t$ to the $(t+1)$-th MCVEM iteration; $\tilde{T}^t$ is defined by: $(\bar{q}, \bar{\psi}) \in \tilde{T}^t(\tilde{q}, \tilde{\psi})$ where $\tilde{\psi} = (\tilde{\theta}, \tilde{\beta})$ iff

$$\begin{aligned}
\bar{q} &\in \operatorname{argmin}_{q\in\mathcal{I}_r} \mathrm{KL}(q; p_{Z|Y}(\cdot|\mathbf{y}; \tilde{\psi})), \\
\bar{\theta} &\in \operatorname{argmax}_{\theta\in\Theta} \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_{Y|Z}(\mathbf{y}|\mathbf{z}; \theta)\ \bar{q}(\mathbf{z}), \\
\bar{\beta} &= \operatorname{argmax}_{b\in\{x, |x-\tilde{\beta}|\leq\gamma^t\}} - \left\{ \sum_{\mathbf{z}\in\mathcal{Z}} H(\mathbf{z}; b)\bar{q}(\mathbf{z}) + \ln \tilde{W}^{J_t, \tilde{\beta}}(b) \right\}.
\end{aligned}$$

$L$ (resp. $\mathcal{L}$) are given by (19) (resp. (20)) and $\mathcal{U} = \mathcal{I}_r \times \Psi$. By Lemma 4, $L$ is a continuous Lyapunov function relative to $(\{T^t\}_t, \mathcal{L})$ and to $(T^*, \mathcal{L})$. Under A3, the level sets of $L$ are compact in $\mathcal{I}_r \times \Psi$. $0 < L(q, \psi) < \infty$ for all $(q, \psi) \in \mathcal{I}_r \times \Psi$ so that $\mathcal{I}_r \times \Psi$ is a denumerable union of the level sets of $L$. Finally, let $(\bar{q}, \bar{\psi}) \in T^t(q, \psi)$ and set $\tilde{q} = \bar{q}$, $\tilde{\theta} = \bar{\theta}$ and $\tilde{\beta} = \operatorname{argmax}_{b\in\{x, |x-\beta|\leq\inf_t \gamma^t\}} \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_Z(\mathbf{z}; b)\bar{q}(\mathbf{z})$. Then $(\tilde{q}, \tilde{\psi}) \in T^*(q, \psi)$ and $L(\bar{q}, \bar{\psi}) \geq L(\tilde{q}, \tilde{\psi})$. Under A4, S2 or S2' holds. This concludes the verification of condition **(a)**. The condition **(b)** results from the condition **(c)** applied with $\mathcal{K} = \{u\}$. We now establish **(c)** and prove that the limit holds $\bar{\mathbb{P}}$-a.s. Hereafter, for $(q^t, \psi^t) \in \mathcal{K}$, let $(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) \in \tilde{T}^t(q^t, \psi^t)$. Define $(\bar{q}^{t+1}, \bar{\psi}^{t+1})$ by $\bar{q}^{t+1} = \tilde{q}^{t+1}$, $\bar{\theta}^{t+1} = \tilde{\theta}^{t+1}$ and

$$\bar{\beta}^{t+1} = \operatorname{argmax}_{\beta\in\{|\beta-\beta^t|\leq\gamma^t\}} \sum_{\mathbf{z}\in\mathcal{Z}} \ln p_Z(\mathbf{z}; \beta)\bar{q}^{t+1}(\mathbf{z}),$$

so that $(\bar{q}^{t+1}, \bar{\psi}^{t+1}) \in T^t(q^t, \psi^t)$. Finally, define $(\check{q}^{t+1}, \check{\psi}^{t+1})$ (resp. $(\hat{q}^{t+1}, \hat{\psi}^{t+1})$) by $\check{q}^{t+1} = \hat{q}^{t+1} = \tilde{q}^{t+1}$, $\check{\theta}^{t+1} = \hat{\theta}^{t+1} = \tilde{\theta}^{t+1}$ and $\check{\beta}^{t+1}$ (resp. $\hat{\beta}^{t+1}$) as $\bar{\beta}^{t+1}$ (resp. $\tilde{\beta}^{t+1}$) by setting $\gamma^t = \infty$.
We establish that for all $\epsilon > 0$,

$$\sum_{t\geq 0} \mathbb{1}_{\{|L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \mathbb{1}_{(q^t, \psi^t)\in\mathcal{K}} \geq \epsilon\}} < \infty \tag{24}$$

$\bar{\mathbb{P}}$-a.s. which is implied, by the second Borel-Cantelli Lemma by the $\bar{\mathbb{P}}$-a.s. convergence of the series with general term

$$\bar{\mathbb{P}}\left( |L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \mathbb{1}_{(q^t, \psi^t)\in\mathcal{K}} \geq \epsilon | \mathcal{F}_t \right);$$

$\mathcal{F}_t$ is the sigma-field $\sigma\left(Z^{j,s}, 0 \le j \le J_s, 1 \le s \le t-1\right)$. With this choice of the point $(\bar{q}^{t+1}, \bar{\psi}^{t+1})$ in the set $T^t(q^t, \psi^t)$, we have

$$|L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| = \left| \sum_{\mathbf{z} \in \mathcal{Z}} \ln \left( \frac{p_Z(\mathbf{z}; \tilde{\beta}^{t+1})}{p_Z(\mathbf{z}; \bar{\beta}^{t+1})} \right) \tilde{q}^{t+1}(\mathbf{z}) \right|$$

$$= \left| \sum_{\mathbf{z} \in \mathcal{Z}} \left\{ H(\mathbf{z}; \tilde{\beta}^{t+1}) - H(\mathbf{z}; \bar{\beta}^{t+1}) \right\} \tilde{q}^{t+1}(\mathbf{z}) + \ln \frac{W(\tilde{\beta}^{t+1})}{W(\bar{\beta}^{t+1})} \right| \qquad (25)$$

**Lemma 10** *Assume A1, A2(i)-(iii). If $\{(q^t, \psi^t)\}_t \in \mathcal{K}$ for some compact $\mathcal{K}$ of $\mathcal{I}_r \times \Psi$, there exist (deterministic) compact sets $\mathcal{C}$ and $\mathcal{C}'$ in $\mathcal{B}$, depending upon $\mathcal{K}$, such that $\{\bar{\beta}^t\}_t \subset \mathcal{C}$ and $\{\check{\beta}^t\}_t \subset \mathcal{C}'$.*

*Proof:* Let $\mathcal{C}''$ be the compact set that contains $\{\beta^t\}_t$. By the implicit function theorem, there exists a continuously differentiable function $\rho : \mathcal{I} \to \mathcal{B}$ such that for all $q \in \mathcal{I}$, $G(q, \rho(q)) = 0$ where $G(q, \beta) = \nabla_\beta \{\sum_{\mathbf{z} \in \mathcal{Z}} \ln p_Z(\mathbf{z}; \beta)\, q(\mathbf{z})\}$. Since $\mathcal{I}$ is compact, the family $\{\check{\beta}^{t+1}\}_t$ is in a compact interval, say $\mathcal{C}'$. Let $\mathcal{C}$ be the smallest closed ball that contains $\bar{\beta}^0 \cup \mathcal{C}' \cup \mathcal{C}''$. We prove that $\bar{\beta}^t \in \mathcal{C}$ for all $t$. $\bar{\beta}^0 \in \mathcal{C}$ by definition of the compact. Let $t \ge 0$; either $\bar{\beta}^{t+1} = \check{\beta}^{t+1}$ and $\bar{\beta}^{t+1} \in \mathcal{C}' \subset \mathcal{C}$, or $\bar{\beta}^{t+1} = \beta^t + \text{sign}(\check{\beta}^{t+1} - \beta^t)\, \gamma^t$; if such, $\beta^t \wedge \check{\beta}^{t+1} \le \bar{\beta}^{t+1} \le \beta^t \vee \check{\beta}^{t+1}$ and $\bar{\beta}^{t+1} \in \mathcal{C}$. □

Let $\text{Cl}(\mathcal{C}_\delta)$ be the closed $\delta$-neighborhood of $\mathcal{C}$, for some positive $\delta$ small enough so that $\text{Cl}(\mathcal{C}_\delta)$ is a compact of $\mathcal{B}$. Under A1 and A2(iii), $H(\mathbf{z}; \cdot)$ and $W$ are continuous for all $\mathbf{z} \in \mathcal{Z}$ and thus, uniformly continuous on $\text{Cl}(\mathcal{C}_\delta)$. Since $\mathcal{Z}$ is finite, there exists $\eta > 0$ depending upon $\text{Cl}(\mathcal{C}_\delta)$, $\epsilon$ and $\delta$ such that for all $(x, y) \in \text{Cl}(\mathcal{C}_\delta)$, $q \in \mathcal{I}_r, \theta \in \Theta$,

$$|x - y| \le \eta \Longrightarrow |L(q, \theta, x) - L(q, \theta, y)| \le \epsilon.$$

Hence,

$$\begin{aligned}
&\bar{\mathbb{P}}\big( |L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} \ge \epsilon \big| \mathcal{F}_t \big) \\
&\quad \le \bar{\mathbb{P}}\big( |L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \ge \epsilon,\, |\bar{\beta}^{t+1} - \tilde{\beta}^{t+1}| < \delta \big| \mathcal{F}_t \big) \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} \\
&\quad + \bar{\mathbb{P}}\big( |L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \ge \epsilon,\, |\bar{\beta}^{t+1} - \tilde{\beta}^{t+1}| \ge \delta \big| \mathcal{F}_t \big) \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} \\
&\qquad \le \bar{\mathbb{P}}\big( |\bar{\beta}^{t+1} - \tilde{\beta}^{t+1}| \ge \eta,\, |\bar{\beta}^{t+1} - \tilde{\beta}^{t+1}| < \delta \big| \mathcal{F}_t \big) \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} \\
&\quad + \bar{\mathbb{P}}\big( |L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \ge \epsilon,\, |\bar{\beta}^{t+1} - \tilde{\beta}^{t+1}| \ge \delta \big| \mathcal{F}_t \big) \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} \\
&\qquad \le 2\, \bar{\mathbb{P}}\big( |\bar{\beta}^{t+1} - \tilde{\beta}^{t+1}| \ge \alpha \big| \mathcal{F}_t \big) \mathbb{1}_{(q^t, \psi^t) \in \mathcal{K}} \qquad (26)
\end{aligned}$$

where $\alpha = \delta \wedge \eta \wedge \inf_t \gamma^t > 0$.

**Lemma 11** *Set $\tilde{G}_{t,b}(q, \beta) = -\nabla_\beta \{\sum_{\mathbf{z} \in \mathcal{Z}} H(\mathbf{z}; \beta) q(\mathbf{z}) + \ln \tilde{W}^{J_t, b}(\beta)\}$. Under A1, A2 and A5, for any $0 < \alpha \le \inf_t \gamma^t$, there exists a deterministic $\iota > 0$ (independent of $t$) such that*

$$|\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha) - G(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha)| \le \iota \;\text{ and }\; |\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} + \alpha) - G(\tilde{q}^{t+1}, \check{\beta}^{t+1} + \alpha)| \le \iota$$

$$\Longrightarrow \quad |\tilde{\beta}^{t+1} - \bar{\beta}^{t+1}| \le \alpha.$$

*Proof:* Let $\rho$ be the continuous function defined in the proof of Lemma 10. For $\alpha > 0$, define $\underline{G} = \min_{q\in\mathcal{I}} G(q, \rho(q) - \alpha)$ and $\bar{G} = \max_{q\in\mathcal{I}} G(q, \rho(q) + \alpha)$. Under A2(iii), $\beta \mapsto G(q, \beta)$ is strictly decreasing for all $q \in \mathcal{I}$, so that $G(q, \rho(q)-\alpha) > 0$ and $G(q, \rho(q)+\alpha) < 0$ for all $q \in \mathcal{I}$. Furthermore, since $q \mapsto G(q, \rho(q)+a)$ is continuous on $\mathcal{I}$ for all $a \in \{-\alpha, \alpha\}$, $\underline{G} > 0$ et $\bar{G} < 0$. Finally, recall that by definition of $\check{\beta}^{t+1}$, $\check{\beta}^{t+1} = \rho(\tilde{q}^{t+1})$. Set $\iota = \frac{1}{2}\min(\underline{G}, -\bar{G})$. Since for all $a \in \{-\alpha, \alpha\}$, $|\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} + a) - G(\tilde{q}^{t+1}, \check{\beta}^{t+1} + a)| \leq \iota$ then $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha) > 0$ and $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} + \alpha) < 0$. We distinguish four cases.

`Case 1:` $\bar{\beta}^{t+1} = \check{\beta}^{t+1}$ and $\tilde{\beta}^{t+1} = \hat{\beta}^{t+1}$. Under A2(iii), $\hat{\beta}^{t+1}$ is the unique solution to $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \cdot) = 0$; hence, $\rho(\tilde{q}^{t+1}) - \alpha \leq \hat{\beta}^{t+1} \leq \rho(\tilde{q}^{t+1}) + \alpha$ and this yields $|\tilde{\beta}^{t+1} - \bar{\beta}^{t+1}| \leq \alpha$.

`Case 2:` $\bar{\beta}^{t+1} \neq \check{\beta}^{t+1}$ and $\tilde{\beta}^{t+1} \neq \hat{\beta}^{t+1}$. We prove that $\bar{\beta}^{t+1} = \tilde{\beta}^{t+1}$. To that goal, we first assume that $\check{\beta}^{t+1} > \beta^t$ which implies, under A2(iii), that $\bar{\beta}^{t+1} = \beta^t + \gamma^t$. If $\check{\beta}^{t+1} - \alpha \geq \beta^t + \gamma^t$, $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \beta) > 0$ for all $\beta^t - \gamma^t \leq \beta \leq \beta^t + \gamma^t$ since under A2(iv), $\beta \to \tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \beta)$ is decreasing. Hence, $\tilde{\beta}^{t+1} = \beta^t + \gamma^t = \bar{\beta}^{t+1}$. If $\check{\beta}^{t+1} - \alpha \leq \beta^t + \gamma^t$, the condition $\alpha \leq \inf_t \gamma^t$ implies that there exists $\beta^t \leq \beta \leq \beta^t + \gamma^t$ such that $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \beta) > 0$; since $\tilde{\beta}^{t+1} \neq \hat{\beta}^{t+1}$, this implies, together with A2(iv), $\tilde{\beta}^{t+1} = \beta^t + \gamma^t$. The case $\check{\beta}^{t+1} < \beta^t$ is along the same lines and is omitted.

`Case 3:` $\bar{\beta}^{t+1} \neq \tilde{\beta}^{t+1}$ and $\tilde{\beta}^{t+1} = \hat{\beta}^{t+1}$. We first assume that $\check{\beta}^{t+1} > \beta^t$ so that $\bar{\beta}^{t+1} = \beta^t + \gamma^t$. Since $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha) > 0$ and $\tilde{\beta}^{t+1} = \hat{\beta}^{t+1}$, $\check{\beta}^{t+1} - \alpha < \tilde{\beta}^{t+1} < \beta^t + \gamma^t$. Furthermore, $\bar{\beta}^{t+1} \neq \check{\beta}^{t+1}$ so that $\bar{\beta}^{t+1} < \check{\beta}^{t+1}$. This yields $|\tilde{\beta}^{t+1} - \bar{\beta}^{t+1}| \leq \alpha$. The case $\check{\beta}^{t+1} < \beta^t$ is along the same lines and is omitted.

`Case 4:` $\bar{\beta}^{t+1} = \check{\beta}^{t+1}$ and $\tilde{\beta}^{t+1} \neq \hat{\beta}^{t+1}$. We first assume that $\check{\beta}^{t+1} > \beta^t$. The conditions $\alpha \leq \inf_t \gamma^t$ and $\bar{\beta}^{t+1} = \check{\beta}^{t+1}$ imply $\beta^t - \gamma^t < \check{\beta}^{t+1} - \alpha \leq \beta^t + \gamma^t$. These inequalities, combined with $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha) > 0$ and $\tilde{\beta}^{t+1} \neq \hat{\beta}^{t+1}$ yield $\tilde{\beta}^{t+1} = \beta^t + \gamma^t$. Finally, $\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} + \alpha) < 0$ and $\tilde{\beta}^{t+1} = \beta^t + \gamma^t$ imply $\beta^t + \gamma^t < \check{\beta}^{t+1} + \alpha$. We thus obtain $|\tilde{\beta}^{t+1} - \bar{\beta}^{t+1}| \leq \alpha$. The case $\check{\beta}^{t+1} < \beta^t$ is along the same lines and is omitted.

□

This yields, on the event $\{(q^t, \psi^t) \in \mathcal{K}\} \in \mathcal{F}_t$,

$$\begin{aligned}
\frac{1}{2}\,\bar{\mathbb{P}}\big(\,|L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \geq \epsilon \big| \mathcal{F}_t\,\big) \\
&\leq \bar{\mathbb{P}}(|\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha) - G(\tilde{q}^{t+1}, \check{\beta}^{t+1} - \alpha)| \geq \iota | \mathcal{F}_t) \\
&\quad + \bar{\mathbb{P}}(|\tilde{G}_{t,\beta^t}(\tilde{q}^{t+1}, \check{\beta}^{t+1} + \alpha) - G(\tilde{q}^{t+1}, \check{\beta}^{t+1} + \alpha)| \geq \iota | \mathcal{F}_t) \\
&\leq \bar{\mathbb{P}}(|\nabla \ln \tilde{W}^{J_t,\beta^t}(\check{\beta}^{t+1} - \alpha) - \nabla \ln W(\check{\beta}^{t+1} - \alpha)| \geq \iota | \mathcal{F}_t) \\
&\quad + \bar{\mathbb{P}}(|\nabla \ln \tilde{W}^{J_t,\beta^t}(\check{\beta}^{t+1} + \alpha) - \nabla \ln W(\check{\beta}^{t+1} + \alpha)| \geq \iota | \mathcal{F}_t). \qquad (27)
\end{aligned}$$

By the Markov's inequality and Lemma 10, this yields for all $r \geq 1$,

$$\begin{aligned}
&\bar{\mathbb{P}}\big(\,|L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi}^{t+1})| \mathbb{1}_{(q^t,\psi^t)\in\mathcal{K}} \geq \epsilon \big| \mathcal{F}_t\,\big) \\
\leq\ & \frac{2}{J_t^r\,\iota^r} \sup_{\beta \in \mathrm{Cl}(\mathcal{C}'_\alpha)} \mathbb{E}_{\lambda,\beta_t}\left[\left|\nabla \ln \tilde{W}^{J_t,\beta^t}(\beta) - \nabla \ln W(\beta)\right|^r\right] \mathbb{1}_{(q^t,\psi^t)\in\mathcal{K}}.
\end{aligned}$$

For $r$ given by A5, there exists a constant $C$ such that for any $t$,

$$\bar{\mathbb{P}}(|L(\tilde{q}^{t+1}, \tilde{\psi}^{t+1}) - L(\bar{q}^{t+1}, \bar{\psi})| \mathbb{1}_{(q^t,\psi^t)\in\mathcal{K}} \geq \epsilon | \mathcal{F}_t) \leq C\, J_t^{-r/2}.$$

The series (24) is thus finite $\bar{\mathbb{P}}$-a.s. for some sequence $\{J_t\}_t$ satisfying A6. This concludes the proof of the first claim (i)(a).

We now check the conditions of Proposition 7. By Lemma 5, $\mathcal{L}$ is a closed set in $\mathcal{I}_r \times \Psi$ so that $\mathcal{L} \cap \mathcal{K}$ is a compact subset of $\mathcal{I}_r \times \Psi$, whatever $\mathcal{K}$ compact. Let $\{u^t = (q^t, \psi^t)\}_t$ be

the stable MCVEM trajectory. It remains to show that the limit holds $\bar{\mathbb{P}}$-a.s. It is sufficient to prove that for any deterministic compact $\mathcal{K} \subset \mathcal{I}_r \times \Psi$,

$$\lim_t |L(\bar{q}^{t+1}, \bar{\psi}^{t+1}) - L(q^{t+1}, \psi^{t+1})| \mathbb{I}_{(q^t,\psi^t)\in\mathcal{K}} = 0 \qquad \bar{\mathbb{P}}\text{-a.s.} \tag{28}$$

for some $(\bar{q}^{t+1}, \bar{\psi}^{t+1}) \in T^t(q^t, \psi^t)$; we choose

$$\bar{q}^{t+1} = q^{t+1}, \quad \bar{\theta}^{t+1} = \theta^{t+1} \quad \bar{\beta}^{t+1} = \text{argmax}_{\beta\in\{|\beta-\beta^t|\le\gamma^t\}}\{\sum_{\mathbf{z}\in\mathcal{Z}} \ln p_Z(\mathbf{z};\beta) q^{t+1}(\mathbf{z})\}.$$

Here again, the limit is a consequence of the $\bar{\mathbb{P}}$-a.s. convergence of the series

$$\sum_{t\ge 0} \bar{\mathbb{P}}\big(|L(\bar{q}^{t+1}, \bar{\psi}^{t+1}) - L(q^{t+1}, \psi^{t+1})| \mathbb{I}_{(q^t,\psi^t)\in\mathcal{K}} \ge \epsilon \big| \mathcal{F}_t\big). \tag{29}$$

Since the number of projections is finite a.s., $(q^{t+1}, \psi^{t+1}) = (\tilde{q}^{t+1}, \tilde{\psi}^{t+1})$ for all $t$ large enough. Thus the series (29) is finite $\bar{\mathbb{P}}$-a.s. if the series (24) is finite $\bar{\mathbb{P}}$-a.s. The convergence of (24) has just been established, thus concluding the proof.

# 10 Application to image segmentation

We show that the model described in Section 5 satisfies the conditions A1 to A6. We assumed that at each pixel, the observations are univariate; this is not at all restrictive and the multi-dimensional case could be considered in the same way.
Conditions **A1**-**A2** are trivially verified; for **A2**, we use the strict concavity of $\beta \mapsto \ln W(\beta)$. Details are omitted.
Regarding assumption **A3**, since $L$ is a continuous positive function and $\mathcal{I}_r$ is bounded it is enough to show that for $q \in \mathcal{I}_r$, $L$ tends to 0 on the boundaries of $\Psi$. Let $\ln L$ be divided in three parts,$\ln L(q,\psi) = a(q,\theta) + b(q,\beta) + c(q)$, where (up to an additive constant independent of the parameters),

$$\begin{aligned} a(q,\theta) &= -\frac{1}{2}\sum_{i\in S}\sum_{k=1}^{K} \left[\ln(\sigma_k) + \sigma_k^{-1}(y_i-\mu_k)^2\right] q_i(e_k), \\ b(q,\beta) &= \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}) \ln\left(p_Z(\mathbf{z};\beta)\right). \end{aligned}$$

For $q \in \mathcal{I}_r$, for all $l = 1, \ldots, K$ there exist $i \in S$ such that $q_i(e_l) > 0$. Then whenever there exists $k$ such that $\sigma_k$ tends to 0 or $\mu_k$ tends to $\pm\infty$, part $\sigma_k^{-1}(y_i - \mu_k)^2$ is the most significant term in expression $a(q,\theta)$. If $\sigma_k$ tends to $+\infty$, then the most significant term in $a(q,\theta)$ is $\ln(\sigma_k)$. In all cases $a(q,\theta)$ tends to 0. When $\beta$ tends to $+\infty$ (resp. $-\infty$) then $p_Z(\mathbf{z};\beta)$ tends to 0 except in $\mathbf{z} \in \arg\max_z h(\mathbf{z})$ (resp. $\mathbf{z} \in \arg\min_z h(\mathbf{z})$) so that clearly $b(q,\beta)$ tends to 0. It follows that **A3** is satisfied.
For **A4**, we show that $\mathcal{L}$ is compact which implies that $L(\mathcal{L})$ is compact since $L$ is continuous. Under the stated assumptions, $\mathcal{L}$ is closed (see Lemma 5) and it remains to prove that $\mathcal{L}$ is bounded. Let us first observe that for $(q^*, \psi^*) \in \mathcal{L}$, $q^*$ is included in a compact set and $\psi^*$ satisfies $\nabla_\psi F(q^*, \psi^*) = 0$, which leads to closed-form expressions

$$\begin{aligned} \forall k, \quad \mu_k^* &= \frac{\sum_{i\in S} y_i q_i^*(e_k)}{\sum_{i\in S} q_i^*(e_k)}, \\ \sigma_k^* &= \frac{\sum_{i\in S}(y_i-\mu_k^*)(y_i-\mu_k^*)^t q_i^*(e_k)}{\sum_{i\in S} q_i^*(e_k)}; \end{aligned}$$

hence, $q^*$ and $\theta^*$ are linked through a continuous and bounded function on $\mathcal{I}_r$ and $\theta^*$ is bounded. By applying the implicit function theorem we prove that the same holds for $\beta^*$ (see the details in the proof of lemma 10 in Appendix 9) which shows that $\mathcal{L}$ is bounded. For **A5**, we can actually show that a more general condition holds: applying the results by [15], we can deduce that the conditions in **A5** hold for all $r \geq 2$ and any initial distribution $\lambda$. Referring to [15, Proposition 1], it is enough to show that for the Markov chain used in the approximation of $W(\beta)$, the state space is *small* (see *e.g.* [26]). The Gibbs sampler is a Markov chain with kernel $P = P_1 P_2 \dots P_N$ where $P_n$ replaces the $n^{th}$ pixel with a draw from the conditional $p_Z(z_n|\mathbf{z}_{S\setminus\{n\}})$ leaving $\mathbf{z}_{S\setminus\{n\}}$ unchanged. Since $\mathcal{Z} = V^N$ is a product space with $V$ finite, the small set property follows easily.

# References


[1] C. Ambroise and G. Govaert. Convergence proof of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters*, 19:919–927, 1998.

[2] J.E.B. Archer and D.M. Titterington. Parameter estimation for hidden markov chains. *Journal of Statistical Planning and Inference*, 108:365–390, 2002.

[3] R.A. Boyles. On the convergence of EM algorithms. *J. Roy. Statist. Soc. Ser. B*, 45(1):47–50, 1983.

[4] W. Byrne and A. Gunawardana. Convergence theorems of Generalized Alternating Minimization Procedures. *Journal of Machine Learning Research*, 1:1–48, 2004.

[5] G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quart.*, 2(1):73–82, 1985.

[6] G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean-field approximations for Markov model-based image segmentation. *Pattern Recognition*, 36:131–144, 2003.

[7] B. Chalmond. An iterative Gibbsian technique for reconstruction of m-array images. *Pattern Recognition*, 22(6):747–761, 1989.

[8] I. Csiszar and G. Tusnady. Information geometry and alternating minimization procedures. *Stat. & Dec.*, (1):205–237, 1984. Supp. Iss.

[9] N. de Freitas, P. Hojen-Sorensen, M. Jordan, and S. Russel. Variational MCMC. Technical report, UC Berkeley, 2001.

[10] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

[12] X. Descombes, R.D. Morris, J. Zerubia, and M. Berthod. Estimation of Markov random field prior parameters using Markov Chain Monte Carlo maximum likelihood. *IEEE Transactions on Image Processing*, 8(7):954–963, 1999.

[13] J.A. Fessler. Comments on "The convergence of mean field procedures for MRF's". *IEEE Transactions on Image Processing*, 7(6):917–, 1998.

[14] F. Forbes and N. Peyrard. Hidden Markov Random Field Model Selection Criteria based on Mean Field-like Approximations. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101, September 2003.

[15] G. Fort and E. Moulines. Convergence of the Monte-Carlo EM for curved exponential families. *Ann. Statist.*, 31(4), 2003.

[16] A. Frigessi, C.-R. Hwang, and L. Younes. Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. *Annals of Applied probability*, 2(3):610–628, 1992.

[17] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

[18] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[19] J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 24:1317–1399, 1989.

[20] C.J. Geyer and E. Thompson. Constrained Monte-Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B*, 54:657–699, 1992.

[21] W.R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in practice*. Chapman and Hall, Suffolk, 1996.

[22] Z. Jordan, M.I. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. 1998.

[23] J.L. Maroquin, S.Mitter, and T. Pogio. Probabilistic solution of ill-posed problems in computer vision. *J. Ann. Statist. Ass.*, 82:76–89, 1987.

[24] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1996.

[25] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

[26] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[27] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variante. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. 1998.

[28] S.F. Nielsen. The stochastic EM: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

[29] B.T. Polyak and A.B. Juditski. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim*, 30(4):838–855, 1992.

[30] W. Qian and D.M. Titterington. Estimation of parameters in hidden markov models. *Philos. Trans. Roy. Soc. London Ser. A*, 337:407–428, 1991.

[31] W. Qian and D.M. Titterington. Discussion of a paper by Geyer and Thompson. *J. Roy. Statist. Soc. Ser. B*, 54:657–699, 1992.

[32] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999.

[33] C. Sminchisescu, M. Welling, and G. Hinton. A Mode-Hopping MCMC sampler. Technical Report CSRG-478, Univ. of Toronto, 2003.

[34] D. Stanford. *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Processes*. PhD thesis, Department of Statistics, University of Washington, Seattle, 1999.

[35] R. Swendsen and J.S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, (58):86–88, 1987.

[36] T. Tanaka. Information geometry of mean-field approximation. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods*, chapter 17. MIT Press, 2001.

[37] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Department of Statistics, 2003.

[38] M.J. Wainwright and M.I. Jordan. A variational principle for graphical models. In S. Haykin, T. Principe, T. Sejnowski, and J. McWhirter, editors, *New Directions in Statistical Signal Processing*, chapter 11. MIT Press, 2005.

[39] G. Wei and M.Tanner. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J.Amer.Statist.Assoc.*, 85:699–704, 1990.

[40] C-H. Wu and P. C. Doerschuk. Cluster Expansions for the Deterministic Computation of Bayesian Estimators Based on Markov Random fields. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 17(3):275–293, 1995.

[41] C.F.J. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.

[42] L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probab. Theory Relat. Fields*, 82:625–645, 1989.

[43] L. Younes. Monte-Carlo maximization of likelihood: A convergence study. Technical report, CMLA, ENS Cachan, France, 2000. Available at http://www.cmla.ens-cachan.fr/Utilisateurs/younes.

[44] W.I. Zangwill. *Nonlinear Programming : a Unified Approach*. Prentice-Hall, 1969.

[45] J. Zhang. The Mean Fields Theory in EM Procedures for Markov Random Fields. *IEEE Transactions on signal processing*, 40(10):2570–2583, 1992.

[46] J. Zhang. The convergence of mean field procedures for MRF's. *IEEE Transactions on Image Processing*, 5(12):1662–1665, 1996.

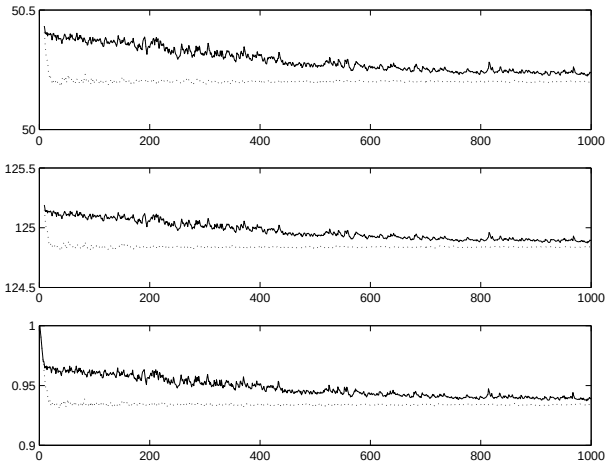


Figure 1: Logo image: MCVEM trajectories when $\mathbf{z}^{0,t} = \mathbf{z}^0$ (solid line) and $\mathbf{z}^{0,t} = \mathbf{z}^{J_{t-1},t-1}$ (dot line). [top] $\mu_1$ (the first 8 values are discarded), [center] $\sigma_1$ (the first 8 values are discarded), [bottom] $\beta$.

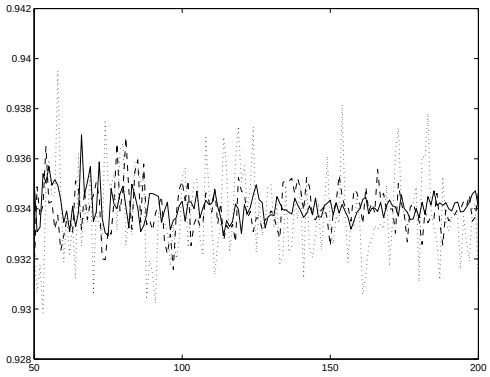


Figure 2: Logo image: $\beta$ trajectory as a function of the number of iterations when $J_t \sim (2t)^{1.01}$ (dot line), $J_t \sim (2t)^{1.3}$ (dash-dot line) and $J_t \sim (2t)^{1.5}$ (solid line). The first 50 iterations are discarded.

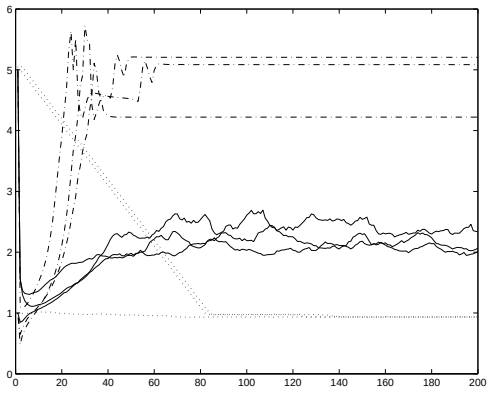


Figure 3: Logo image: $\beta$ trajectory versus the number of iterations for different parameter starting values, with *Mean Field* (dot line), *Simulated Field* (solid line) and *MCVEM* (dash-dot line)

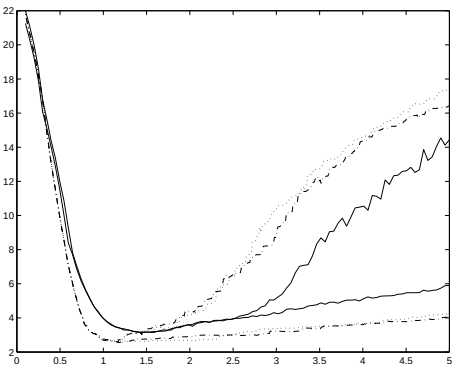


Figure 4: Logo image : Error rate versus $\beta$ obtained by *Mean Field* (dot line), *Simulated Field* (solid line) and *MVCEM* (dash-dot line), when the segmentation algorithm is started from two different initial classifications.

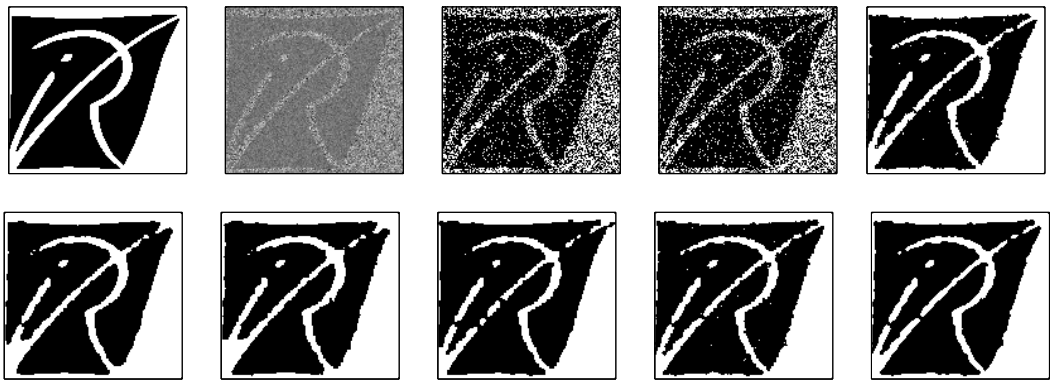

Figure 5: Logo image: [top, from left to right] original image, noise-corrupted image, initial segmentation using kmeans, ind-EM, MC2-EM; [bottom, from left to right] Gibbsian-EM, Simulated Field, Mean Field, MCVEM, MCVEM + Median Filter

.

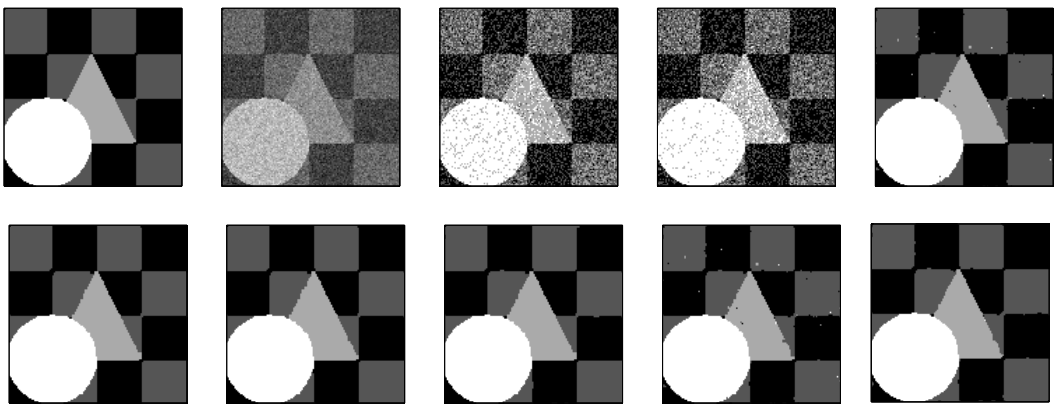

Figure 6: Triangle image: [top, from left to right] original image, noise-corrupted image, initial segmentation using kmeans, ind-EM, MC2-EM; [bottom, from left to right] Gibbsian-EM, Simulated Field, Mean Field, MCVEM, MCVEM + Median Filter

.

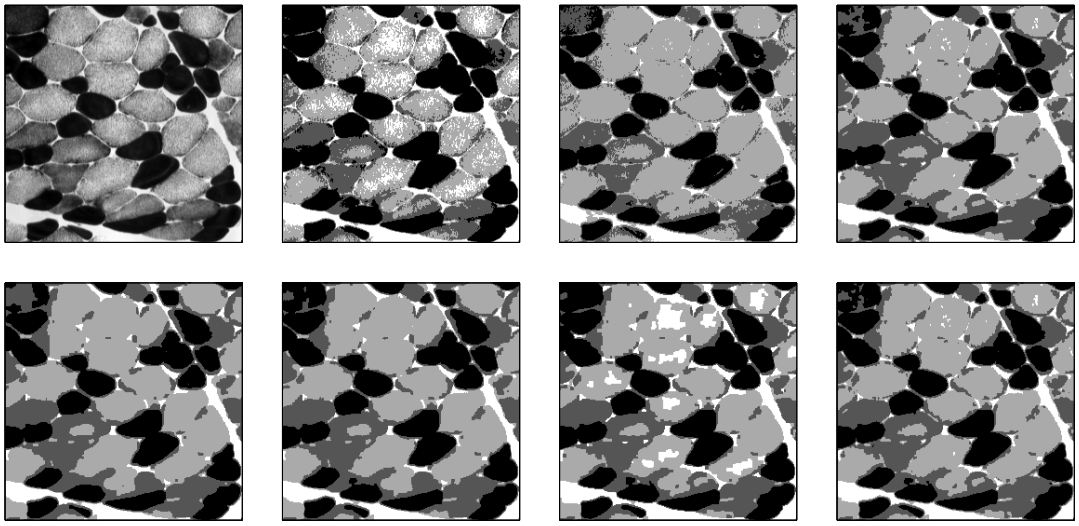

Figure 7: Muscle image: [top, from left to right] original image, $k$-means, *ind-EM*, *MC2-EM*; [bottom, from left to right] *Gibbsian-EM*, *Simulated Field*, *Mean Field*, *MCVEM*.

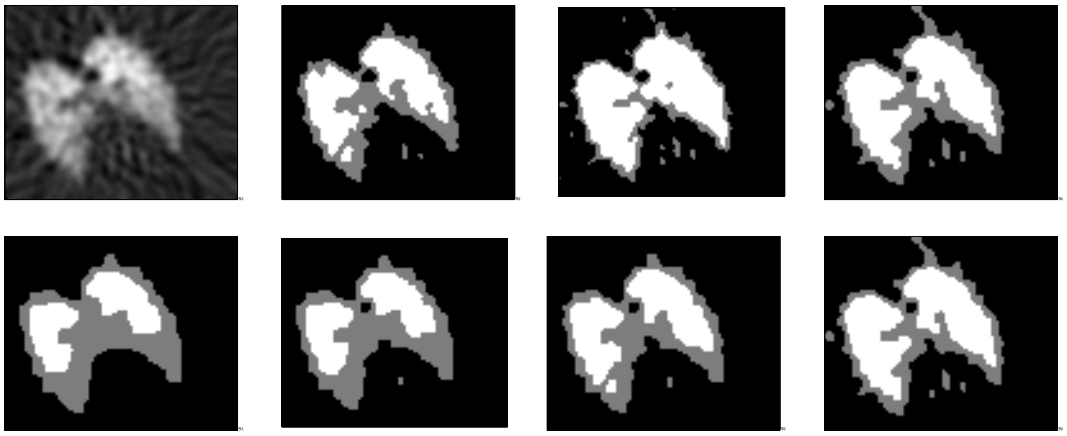

Figure 8: PET image of a dog lung: [top, from left to right] original image, initial segmentation, *ind-EM*, *MC2-EM*; [bottom, from left to right] *Gibbsian-EM*, *Simulated Field*, *Mean Field*, *MCVEM*.

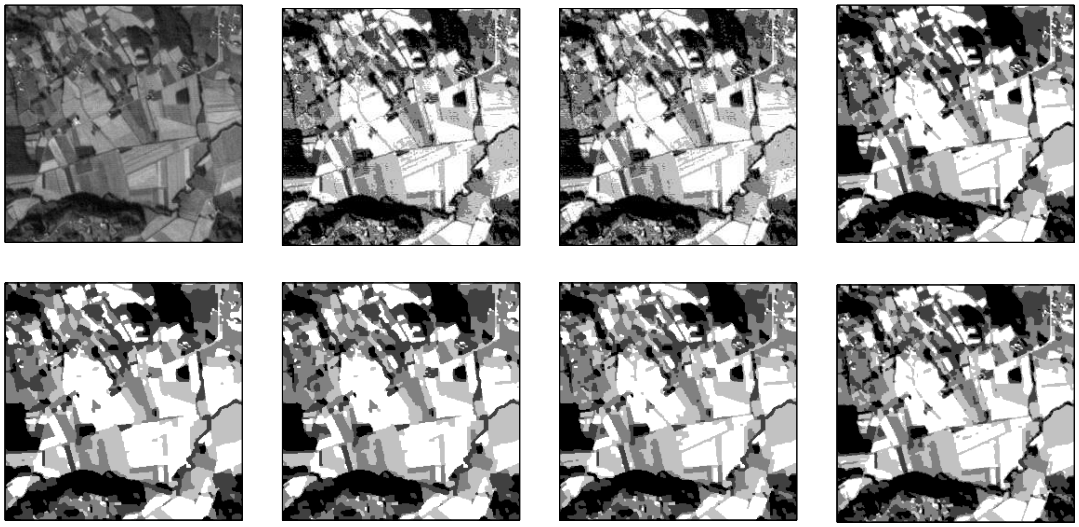

Figure 9: Satellite image: [top, from left to right] original image, initial segmentation, *ind-EM*, *MC2-EM*; [bottom, from left to right] *Gibbsian-EM*, *Simulated Field*, *Mean Field*, *MCVEM*.

| algorithm | $\beta$ | $\mu_1$ | $\mu_2$ | $\sigma_1$ | $\sigma_2$ | error rate | ref. |
|---|---|---|---|---|---|---|---|
| true values | 0.78 | 1 | 2 | 0.5 | 0.5 | - | - |
| ind-EM | - | 1.01 | 2.01 | 0.50 | 0.50 | 15.91 | 15.85 |
| | - | (3.16 10e-2) | (2.65 10e-2) | (1.41 10e-2) | (1.26 10e-2) | (0.33) | (0.26) |
| Mean Field | 0.94 | 1.01 | 2.00 | 0.51 | 0.50 | 10.28 | 9.77 |
| | (2.83 10e-2) | (1.73 10e-2) | (1.41 10e-2) | (10e-2) | (10e-2) | (0.49) | (0.42) |
| Simulated Field | 0.78 | 1.00 | 2.00 | 0.50 | 0.50 | 10.96 | 11.04 |
| | (2.24 10e-2) | (10e-2) | (10e-2) | (10e-2) | (10e-2) | (0.43) | (0.48) |
| MCVEM | 0.73 | 0.98 | 2.02 | 0.48 | 0.48 | 9.87 | 9.77 |
| | (1.77 10e-2) | (1.13 10e-2) | (1.12 10e-2) | (7.3 10e-3) | (7.1 10e-3) | (0.42) | (0.42) |
| MC2-EM | 0.77 | 1.00 | 2.00 | 0.50 | 0.50 | 9.81 | 9.81 |
| | (1.44 10e-2) | (1.19 10e-2) | (1.20 10e-2) | (0.80 10e-2) | (0.81 10e-2) | (0.39) | (0.39) |
| Gibbsian-EM | 0.77 | 1.00 | 2.00 | 0.50 | 0.50 | 9.79 | 9.81 |
| | (2.23 10e-2) | (1.14 10e-2) | (1.24 10e-2) | (0.82 10e-2) | (0.84 10e-2) | (0.40) | (0.39) |

Table 1: Parameter estimates and error rates for the hidden 2-color Potts model with $\beta = 0.78$ (first order neighborhood). The results are mean values over 20 runs, the standard deviations are also reported in parenthesis.

| algorithm | $\beta$ | error rate | ref. |
|---|---|---|---|
| true values | 0.90 | - | - |
| ind-EM | - | 21.31 (0.60) | 21.14 (0.50) |
| Mean Field | 1.03 (2.45 10e-2) | 14.03 (0.60) | 13.78 (0.59) |
| Simulated Field | 0.90 (2.45 10e-2) | 15.67 (0.56) | 15.69 (0.64) |
| MCVEM | 0.85 (1.89 10e-2) | 14.02 (0.59) | 13.78 (0.59) |
| MC2-EM | 0.89 (1.36 10e-2) | 13.77 (0.53) | 13.79 (0.54) |
| Gibbsian-EM | 0.89 (2.23 10e-2) | 13.77 (0.53) | 13.79 (0.54) |

Table 2: Beta estimates and error rates for the hidden 3-color Potts model with $\beta = 0.9$ (first order neighborhood). The results are mean values over 20 runs, the standard deviations are also reported in parenthesis.

| algorithm | $\beta$ | error rate | ref. |
|---|---|---|---|
| true values | 1 | - | - |
| ind-EM | - | 24.23 (0.54) | 23.87 (0.45) |
| Mean Field | 1.05 ( 1.95 10e-2) | 18.32 (0.51) | 18.38 (0.45) |
| Simulated Field | 0.90 (1.64 10e-2) | 20.73 (0.55) | 20.82 (0.48) |
| MCVEM | 0.81 (1.17 10e-2) | 18.66 (0.50) | 18.38 (0.45) |
| MC2-EM | 0.89 (1.07 10e-2) | 18.15 (0.49) | 18.24 (0.47) |
| Gibbsian-EM | 0.89 (1.67 10e-2) | 18.14 (0.50) | 18.24 (0.47) |

Table 3: Beta estimates and error rates for the 4-color Potts model with $\beta = 1$ (first order neighborhood). The results are mean values over 20 runs, the standard deviations are also reported in parenthesis.

| algorithm | $\beta$ | $\mu_1$ | $\mu_2$ | $\sigma_1$ | $\sigma_2$ | error rate |
|---|---|---|---|---|---|---|
| true values | - | 51 | 255 | 130 | 300 | - |
| ind-EM | - | 52 | 255 | 128 | 304 | 22.69 |
| Mean Field | 4.22 | 53 | 260 | 130 | 306 | 3.10 |
| Simulated Field | 2.15 | 52 | 250 | 128 | 302 | 3.42 |
| MCVEM | 0.93 | 50 | 262 | 125 | 305 | 2.89 |
| MC2-EM | 0.91 | 50 | 261 | 125 | 305 | 2.89 |
| Gibbsian-EM | 1.82 | 52 | 251 | 128 | 303 | 2.92 |

Table 4: Parameter estimates and error rates for the degraded 2-color logo image.

| algorithm | $\beta$ | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | error rate |
|---|---|---|---|---|---|---|---|---|---|---|
| true values | - | 1 | 2 | 3 | 4 | 0.5 | 0.5 | 0.5 | 0.5 | - |
| ind-EM | - | 0.85 | 1.69 | 2.54 | 3.93 | 0.44 | 0.42 | 0.46 | 0.53 | 29.4 |
| Mean Field | 3.99 | 0.99 | 2.00 | 2.98 | 4.01 | 0.49 | 0.50 | 0.48 | 0.50 | 0.44 |
| Simulated Field | 3.46 | 1.00 | 2.00 | 2.97 | 4.01 | 0.49 | 0.50 | 0.48 | 0.50 | 0.40 |
| MCVEM | 1.27 | 0.99 | 2.00 | 2.98 | 4.01 | 0.48 | 0.48 | 0.46 | 0.50 | 0.80 |
| MC2-EM | 1.26 | 0.99 | 2.00 | 2.97 | 4.01 | 0.48 | 0.48 | 0.46 | 0.49 | 0.81 |
| Gibbsian-EM | 3.01 | 1.00 | 2.00 | 2.98 | 4.00 | 0.49 | 0.50 | 0.48 | 0.50 | 0.31 |

Table 5: Parameter estimates and error rates for the degraded 4-color image.

Unité de recherche INRIA Rhône-Alpes

655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes

4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique

615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur

INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)

http://www.inria.fr

ISSN 0249-6399